MEDLINE METRIC: A METHOD TO ASSESS MEDICAL STUDENTS' MEDLINE SEARCH EFFECTIVENESS

Gale G. Hannigan, MLS, MPH

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2000

APPROVED:

Ana D. Cleveland, Major Professor
Rachel Bramson, Committee Member
Donald B. Cleveland, Committee Member
William K. Brookshire, Committee Member
Lawrence R. Wheeless, Committee Member
Philip M. Turner, Director of the Interdisciplinary Program in Information Science
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Hannigan, Gale G., MEDLINE Metric: A method to assess medical students' MEDLINE search effectiveness. Doctor of Philosophy (Information Science), May 2000, 109 pp., 13 tables, 10 figures, references, 90 titles.

Medical educators advocate the need for medical students to acquire information management skills, including the ability to search the MEDLINE database. There has been no published validated method available to use for assessing medical students' MEDLINE information retrieval skills.

This research proposes and evaluates a method, designed as the MEDLINE Metric, for assessing medical students' search skills. MEDLINE Metric consists of: (a) the development, by experts, of realistic clinical scenarios that include highly constructed search questions designed to test defined search skills; (b) timed tasks (searches) completed by subjects; (c) the evaluation of search results; and (d) instructive feedback. A goal is to offer medical educators a valid, reliable, and feasible way to judge mastery of information searching skill by measuring results (search retrieval) rather than process (search behavior) or cognition (knowledge about searching).

Following a documented procedure for test development, search specialists and medical content experts formulated six clinical search scenarios and questions. One hundred and forty-five subjects completed the six-item test under timed conditions. Subjects represented a wide range of MEDLINE search expertise. One hundred twenty complete cases were used, representing 53 second-year medical students (44%), 47 fourth-year medical students (39%), and 20 medical librarians (17%). Data related to educational level, search training, search experience, confidence in retrieval, difficulty of search, and score were analyzed.

Evidence supporting the validity of the method includes the agreement by experts about the skills and knowledge necessary to successfully retrieve information relevant to a clinical question from the MEDLINE database. Also, the test discriminated among different performance levels. There were statistically significant, positive relationships between test score and level of education, self-reported previous MEDLINE training, and self-reported previous search experience.

The findings from this study suggest that MEDLINE Metric is a valid method for constructing and administering a performance-based test to identify mastery in searching the MEDLINE database. The test's reliability needs to be established.

ACKNOWLEDGMENTS

This research was funded, in part, by the Institute for Scientific Information/ Medical Library Association Doctoral Fellowship awarded in 1996. The author also wishes to acknowledge search specialists C. Foster, B. Henry, A. McKibbon, M. Vugrin and medical content experts S. Bartold, R. Bramson, D. FitzSimon-Williams, and S. Smith for their involvement in developing the test. R. Gray, K. Saenz, J. Tonn-Bessent and A. Zaragoza assisted organizing test materials, administering the test and/or coding results. C. Henderson gave advice about data analysis and L. Lotman provided document processing and editorial support. Their efforts are greatly appreciated. I am particularly appreciative of the time my committee members spent working with me (R. Bramson, W. Brookshire, D. Cleveland, and L. Wheeless). I wish especially to acknowledge the support of my committee chair and advisor, Dr. Ana D. Cleveland.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

To practice medicine in the 21st century, medical students educated in the 20th century must be given strong grounding in the use of computer technology, to manage information, support patient care decisions, select treatments, and develop their abilities as lifelong learners [1].

This study describes and evaluates a method for assessing medical students' proficiency in using the computerized medical literature information retrieval system MEDLINE.

Background of the Study

The "explosion" in the number of publications in the biomedical literature makes it impossible for physicians to rely on routine review of selected journals to meet their information needs. Former *Journal of the American Medical Association* (JAMA) editor George Lundberg noted:

> [I]f a physician were to attempt to keep up with this literature by reading two articles each day, at the end of one year, that physician would be more than sixty centuries behind. If physicians were to read everything of possible biomedical relevance, they would need to read about 6,000 articles a day. [2]

Of course, no one expects physicians to read all, or even a large part, of the biomedical literature. One does assume that they should be adept at using the medical literature to find information relevant to a clinical question or problem.

An important step in doing this involves searching MEDLINE, the U.S. National Library of Medicine's (NLM) computerized database of citations and abstracts of articles from approximately 4,300 biomedical journals. By constructing an appropriate search strategy, it is possible to limit retrieval from the more than 10 million citations in the database to a small number of citations relevant to the topic searched.

The way in which physicians access MEDLINE has changed since the database was first available nearly 30 years ago. Then, access to MEDLINE was through the medical library and mediated by a librarian who had special training in searching the database and an account with the NLM. Libraries typically charged a fee for these mediated searches. Medical students rarely requested MEDLINE searches; they used the printed counterpart, Index Medicus, instead. In the 1980s, with the advent of microcomputers and CD-ROM technology, commercial companies and the NLM developed "end user" search software so that nonexperts could directly search databases. Librarians shifted their efforts from providing mediated searches to training end users to search on their own. Today, there is free access to MEDLINE through various services on the Internet, as well as commercial MEDLINE products. Medical school libraries support access to and provide training for at least one of the various MEDLINE search systems for end users and continue to offer mediated searches.

Also, during the past 20 years, several published reports have advocated reform in medical education [3-6]. Among the key recommendations is a call to limit the amount of factual information students are expected to memorize and to equally emphasize the acquisition and development of important skills, values, and attitudes for

the practice of medicine. These reports encourage medical schools to prepare students to learn throughout their professional lives by fostering the development of problem-solving, self-directed learning, and information management skills. They recommend that faculty use evaluation methods appropriate for judging analytical and problem-solving abilities [7]. Computerized information retrieval is among those skills specifically mentioned.

The Association of American Medical Colleges (AAMC) sponsored a report published in 1993 that reviewed medical schools' progress toward achieving the goals set by previous calls for reform [8]. Known as the ACME-TRI Report, it documents the results of a survey of the 143 allopathic medical colleges in the United States and Canada. Medical school deans were asked to address 12 topics (based on 18 recommendations) and describe constraints, opportunities, and difficulties in implementing change. Eighty-four schools responded. Three of the topics are relevant to this dissertation: (1) specify what students should learn and the skills and attitudes they should develop, (2) foster self-directed learning and lifelong learning skills, and (3) develop information management skills [9]. Results of the survey indicate that schools have had difficulty defining educational goals, that faculty are reluctant to switch from being transmitters of information to facilitators of learning, that several schools equate self-directed learning with computerized literature search and retrieval, and that there has been "considerable effort in providing students with skills to do on-line literature searches" [10].

The AAMC followed up on the ACME-TRI Report with its Medical School Objectives Project (MSOP) to "develop a consensus within the medical education community on the attributes that medical students should possess at the time of graduation" [11]. Among the stated objectives is "the ability to retrieve (from electronic databases and other resources), manage, and utilize biomedical information for solving problems and making decisions that are relevant to the care of individuals and populations" [12]. Subsequently, the AAMC developed educational objectives specifically for medical informatics. Medical informatics is defined as "the rapidly developing scientific field that deals with the storage, retrieval, and optimal use of biomedical information, data, and knowledge for problem solving and decision making" [13]. The MSOP curriculum for informatics specifies that students should demonstrate the ability to perform database searches and refine search strategies to improve retrieval [14].

Statement of the Problem

The medical education community has indicated the need for medical students to acquire information management skills, including the ability to search the MEDLINE database. At this time, there is no published validated method available to use for assessing medical students' MEDLINE information retrieval skills.

Purpose of the Study

This dissertation describes and evaluates a method to assess medical students' ability to search MEDLINE to find literature highly relevant to patient care questions.

Significance of the Study

By all indications, the MEDLINE database will continue to be used by medical students and practitioners in order to locate information relevant to research and patient care problems. Their success will depend, in part, on the contents of the database and the effectiveness of the search software and, in part, on their search skills. Outside of a formal teaching environment, database searching is typically a private activity with no external feedback. Unless users are made aware of the relevant citations they miss, they may be satisfied with less-than-optimal search results.

Many efforts to assess search performance evaluate the search process and are based on examining the user's printed search strategy or recorded transaction logs. Modern search systems are designed to assist users in the formulation of a search strategy, for example, by mapping to a controlled vocabulary, by presenting applicable subheadings, or by enabling the user to search for relevance-matched results. It is no longer obvious from the user's printed search strategy what was intended by the user and what assistance the system provided. This inability to distinguish the intentions of the user from the actions of the system makes the evaluation of search skill based on search strategy more difficult. Also, the analysis of detailed transaction logs can be quite time-consuming and complex; search logs may not be readily available. In any case, search strategy may not be the most appropriate measure. Search strategy illustrates

technique, but experienced searchers know that there are many good ways to formulate a search. The ultimate goal is to locate the most relevant articles efficiently.

The proposed method of assessment uses search results rather than the search strategy to measure search effectiveness. The evaluation of the outcomes, rather than process, is a more meaningful measure of what may, in fact, occur outside of the testing environment. An evaluation method that scores results is easier and quicker to grade, making it more feasible to use. A results-based assessment is less affected by variations in search style and should be transferable across different search systems, assuming that those systems provide similar functionality.

A valid criterion-based test (i.e., one that "covers a narrow domain and is used for mastery decisions" [15]) would identify students who need remediation in search skills. It would provide an objective indication of performance level, and make available to medical schools one of the first medical informatics evaluation tools. A reliable method to assess search performance would also provide feedback and direction for the content of MEDLINE skills curricula.

Research Question*s*

1. Is the proposed method for assessing search effectiveness valid?
2. Does the test discriminate among different performance levels?
3. Are the results of the test reliable?
4. How many test searches are needed to reasonably assess a student's performance level?

5. Is the method of assessment reliable across different MEDLINE search systems?

6. What is the relationship between level of education and test score?

7. What is the relationship between previous MEDLINE search training and test score?

8. What is the relationship between previous MEDLINE search experience and test score?

Assumptions

Assumption 1. Questions from clinical scenarios can be made explicit enough to result in agreement on the key citations.

Previous research suggests that "different searchers for the same question see and interpret different things in a question, represent them by different linguistic and/or logical constructs, and retrieve different things from a file" [16]. McKibbon found that there was only 25% overlap in retrieval of relevant citations by expert searchers, which gives one pause about the ability of searchers to identify the same most relevant citations, or even the feasibility of assessing search skill based on specific retrieval. In the real world, searchers do not have to get the same results to retrieve the same information. As she points out, "there is much redundancy in the medical literature but it is not known how well this compensates disparate search results" [17]. Her study was based on spontaneous clinical questions, not on questions constructed to test search skills. This study uses carefully constructed questions, each designed and tested at several levels to have an identified, agreed-upon subset of citations that are considered

definitely relevant. The test questions may be somewhat artificial compared to spontaneous clinical questions. It is assumed that the test search questions are archetypes of spontaneous clinical questions and that success in finding the key citations translates into success in finding highly relevant citations for real-life questions.

Assumption 2. Each of the search questions is equally difficult and can be completed in 20 minutes.

The same amount of time was allotted for each search and, for the total score, each question is weighted equally. The test was timed, with 20 minutes for each of the six searches. There are several reasons to time the test. The comparison of students' scores will be more valid if subjects are allotted the same amount of time. Twenty minutes is a reasonable length of time in which to complete a clinical search. The literature suggests that the time to complete an "average" search ranges from 6 to 37 minutes. This author previously administered tests of the search process using 15-minute intervals; most students were able to complete their searches in that time. The time constraint also encourages the subjects to be efficient—they will have to use search techniques to narrow their retrieval because they will not have unlimited search time to sift through many citations. An important element of search effectiveness is the ability to retrieve relevant citations quickly. Inefficiency detracts from the usefulness of searching as a practical tool for information retrieval.

Assumption 3. The subjects are representative of the more general population, and the groups represent different levels of search skill.

Test subjects were recruited from students affiliated with Texas A&M

University and librarians from U.S. medical school libraries. The sample was stratified to include biomedical science undergraduates, medical students, and librarians. It is assumed that this collection of individuals is representative of different levels of search skills. Another assumption is that Texas A&M fourth-year medical students are representative of the population of fourth-year medical students in general, the target group for this test. There is reason to believe that this is the case. Texas A&M medical students perform comparably to other medical students in the US on standardized tests (MCAT, USMLE) and in competition for residency positions.

Study Limitations

Limitation 1: The number of key citations varies with each question.

The most explicit search question is a question for which there is one and only one citation that is relevant; for example: "In the 1980s, Franz Ingelfinger, editor of the *New England Journal of Medicine* wrote an editorial about arrogance. Find the citation to this editorial." Only one record in the MEDLINE database is relevant for this search and its relevancy is undisputed—it is the citation to the editorial. A question such as this is not trivial for assessing search skills, nor is it an artificial question. Many times, however, practitioners are looking for a set of relevant articles about a disease or treatment. The number of articles in the set depends on the available literature on that topic. Most of the searches in the study were not for known items, but for items most relevant to a given question about a disease or condition.

Limitation 2: The MEDLINE database changes over time.

Since the MEDLINE database is updated every two weeks, new citations

relevant to the search may be added at any time. In order to secure the same key citations over time, question may need to include qualifiers to limit the search by date (e.g., “Find articles from 1994 to 1996 most relevant to this question”).

Limitation 3: The sample size is small.

Despite several efforts to recruit students to take the test, medical students are extremely busy and generally not interested in making a two-hour time commitment, even with reimbursement.

Limitation 4: Instructional settings for the test differed.

The second-year medical students took the test in two large groups, as part of a Preceptorship course workshop. The undergraduates and third and fourth-year medical students scheduled time with a test administrator. The librarians self-administered the test. Although the testing environments differed, the instructions and test materials were the same for all subjects.

CHAPTER 2

REVIEW OF THE LITERATURE

Physicians need efficient access to current information. Studies show that there may be considerable lag time between the publication of information about an important medical intervention and its widespread adoption [18]. Standard practice, and even the opinion of experts, may not reflect current research findings [19]. Keeping up with medical information is a serious challenge for physicians. This is due to lack of time [20], and the lack of information management skills and readily available resources [21]. Studies also indicate that the medical literature itself needs improvement in order to be more useful to physicians. This literature is replete with small studies of questionable methodology reporting contradictory results and sometimes lacking in clinically useful information [22,23]. The increasing use of structured abstracts [24], meta-analyses of clinical trials data and the dissemination of resulting systematic reviews [25], and the development of practice guidelines [26] represent efforts to organize medical research information to make it more explicit and easier to assimilate into clinical practice.

An ongoing series in *The Journal of the American Medical Association* (JAMA) [27] provides tutorials for developing physicians' skills to practice what is called "evidence-based medicine." Evidence-based medicine has been defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [28]. Bibliographic information retrieval systems,

such as those developed for the MEDLINE database, enable users to quickly locate subject-specific literature from among millions of biomedical journal articles indexed by the database. Several articles in the JAMA series, "Users' Guides to the Medical Literature," recommend using MEDLINE as a beginning step. The authors of the first article in the series state that "clinicians can easily acquire the basic skills and learn to retrieve the same number of relevant citations as librarians, even if their searches remain a bit messier" [29].

MEDLINE

MEDLINE is the premier bibliographic database for biomedical research and clinical sciences. The database is produced by the U.S. National Library of Medicine and contains more than 10 million records dating back to 1966. Approximately 4,300 biomedical journals are indexed in the MEDLINE database, and 7,300 new citations are added each week [30].

The evolution of MEDLINE describes the development of one of the earliest informatics projects in medicine. In 1879, Dr. John Shaw Billings, a physician assigned to the U.S. Army Surgeon General's Office, published the first issue of *Index Medicus*:

> [to] record the titles of all new publications in Medicine, Surgery, and the collateral branches, received during the preceding month. . . . The periodicals thus indexed will comprise all current medical journals and transactions of value, so far as they can be obtained. At the close of each yearly volume a double index of authors and subjects will be added, forming a complete bibliography of medicine for the preceding year. [31]

Billings was a forward-thinking person; he predicted that the publication "may expand beyond anything now promised" [32]. In fact, his index grew along with the medical literature—from 20,169 references in 1879 to more than 230,000 references in 1979.

In the 1960s, the National Library of Medicine (NLM) began using mainframe computers to produce *Index Medicus*. A byproduct of this effort was MEDLINE, a computerized version of the index. In 1971, with the development of search retrieval software, MEDLINE became one of the first online bibliographic databases. NLM limited MEDLINE access to people who attended training sessions, typically medical librarians who became "mediators" between the database and the researchers and clinicians who wanted to conduct a search. Searchers used computer terminals and 300-baud modems to connect to the NLM database in Bethesda, Maryland. Many medical school libraries charged a fee for these mediated searches. People who did not want to pay, including most students, continued to use the printed *Index Medicus*.

Since its beginning, MEDLINE's database records were highly structured (see Figure 1). This structure enables searchers to limit retrieval by several fields (e.g., publication type, language, subject). NLM indexers use a controlled vocabulary, known as MeSH (Medical Subject Headings), to index articles. A controlled vocabulary relates articles on similar topics by using the same indexing term, even when the articles themselves use different terms. For example, the MeSH term "Kidney Neoplasms" is used for articles with titles like "Renal Carcinoma," "Kidney Cancer," or "Renal Cell Adenocarcinoma." A searcher who knows to use the MeSH term will retrieve all of those articles. A searcher who relies on words that are not MeSH terms will only retrieve articles that use exactly those textwords.

MEDLINE is a more powerful tool for medical information retrieval than *Index Medicus.* The database record has more access points than are feasible to use in the printed index, it is easy to search for combined concepts (e.g., a particular drug with a

specific disease) and, since 1980, author abstracts are included for the majority of the articles indexed.

Unique Identifier
20001201
Authors
Ravaud A. Debled M.
Institution
Department of Medicine, Institut Bergonie, Regional Cancer Centre, Bordeaux, France. ravaud@bergonie.org
Title
Present achievements in the medical treatment of metastatic renal cell carcinoma. [Review] [62 refs]
Source
Critical Reviews in Oncology-Hematology. 31(1):77-87, 1999 Jun.
NLM Journal Code
ago
Country of Publication
Ireland
MeSH Subject Headings
Antineoplastic Agents / tu [Therapeutic Use]
*Carcinoma, Renal Cell / sc [Secondary]
*Carcinoma, Renal Cell / th [Therapy]
Human
Immunotherapy
Interferon-alpha / tu [Therapeutic Use]
Interleukin-2 / tu [Therapeutic Use]
*Kidney Neoplasms / se [Secretion]
*Kidney Neoplasms / th [Therapy]
Registry Numbers
0 (Antineoplastic Agents). 0 (Interferon-alpha). 0 (Interleukin-2).
ISSN
1040-8428
Publication Type
Journal Article. Review. Review, Tutorial.
Language
English
Entry Month
200001. Entry Week: 2000011.

Figure. 1. Sample MEDLINE record.

During the 1980s, MEDLINE became available through commercial vendors. They developed more user-friendly search systems and delivered MEDLINE not only online but also in CD-ROM format. NLM developed its own end-user search system, *Grateful Med*, and promoted its use by physicians. Many librarians started teaching end-users search techniques [33], and saw a decline of mediated searches [34] as more and more people began to do their own searching. In 1990, Haynes et al. identified six vendors offering 13 online MEDLINE products and seven vendors offering 14 different CD-ROM MEDLINE formats [35]. By the 1990s, most medical school libraries subsidized access to a MEDLINE search system designed for end-users [36] and offered user search training. In 1997, NLM announced free access to MEDLINE on the World Wide Web through two end-user systems: Internet Grateful Med and PubMed.

Studies Related to the Value of MEDLINE

MEDLINE has been the subject of several studies, many of which are important to this research. A few address the value or impact of MEDLINE in the practice of medicine, indirectly answering the question: "Why should medical students become competent MEDLINE searchers?"

There is evidence that MEDLINE is a useful tool for locating information to answer clinical questions. To assess the impact of MEDLINE on decision making and patient care, the National Library of Medicine conducted a study using the Critical Incident Technique (CIT) [37,38]. Five hundred and forty-five health professionals who used MEDLINE (either directly or through the services of a librarian) provided 1,158 Critical Incident Reports based on their assessment of the results of those searches. Respondents characterized 86% of the incidents as having an impact on professional

activities. Approximately three out of five respondents identified as least one search as having high impact on patient care. A detailed analysis of 26 searches characterized as ineffective suggested that the searchers did not make effective use of the MeSH vocabulary.

Chambliss and Conley [39] studied 86 questions asked by family medicine physicians. Fifty-four percent were "fully or nearly fully answered," and 71% of these were answered using MEDLINE. In another study, Scura and Davidoff [40] followed up on 50 information requests made by resident physicians to determine the value of MEDLINE searches to patient care. Twenty percent indicated that the search directly influenced patient management, either in the diagnostic workup or for therapy. Based on the cost of a search, the authors speculated that the cost effectiveness of literature searching might be comparable to information from laboratory studies.

Klein et al. [41] conducted a study at three Detroit hospitals to evaluate the relationship between mediated (by a librarian) MEDLINE literature searches and patients' length of stay and expenses. They compared the hospital costs and length of stay of 192 test cases, for which MEDLINE searches were requested, with 10,409 control cases—those for whom no search was requested. The test cases represented people who had more severe illness. Data analysis showed that the ratio of costs (total dollar cost for stay/average control cost for that diagnosis-related group) is lower when a search is done earlier in the patient's stay. The authors propose that the MEDLINE searches provided information that prompted a different diagnosis or therapy.

Studies Related to User Performance

Other studies relevant to this research analyze user behavior and the effect of certain factors such as cost or training. In these studies, researchers often define performance standards and methods of measurement; some use the quality of searches as an outcome measure. The focus of the review will be on studies that describe methods for assessing user performance and the quality of a search. Most of the articles are about research that is specific to MEDLINE, although a few more general major studies and reviews are also included.

First, some definitions are in order. Common measures in information retrieval systems research are relevance, precision, recall, and specificity. In their book, *Measurement in Information Science*, Boyce, Meadow, and Kraft [42] note that relevance "measures the relationship between a question and an answer." They identify two general meanings of the term "relevance." One is relevance as related to a topic and the other is relevance as utility. While there is usually agreement among individuals about the relatedness of a retrieved citation, the utility of a citation is highly dependent on the person with the information need. For example, suppose two physicians search for information about the drug adriamycin and retrieve an article about the cardiotoxic effects of that drug. Both would probably agree that the article is relevant in terms of being related to the topic; however, the physician who was not previously aware of the drug's cardiotoxic effects might rate the article more useful than would the physician who already knew that information. In this case, relevance is a more subjective measure, dependent on the individual making the evaluation and that person's previous knowledge.

Precision and recall are often used to measure information retrieval system performance, but these measures may be used in examining user performance as well. The definitions of precision and recall depend on the assumption that an item is either relevant or not—a binary sort. A two-by-two contingency table (Table 1) helps clarify the formulas [43]:

Table 1

Retrieval Precision and Recall

	Relevant	Not Relevant
Retreived	a	b
Not Retrieved	c	d

Precision is the ratio of the number of relevant citations retrieved to the total number of citations retrieved, or a/a+b. Recall is the ratio of the number of relevant citations retrieved to the number of relevant citations in the database, or a/a+c. Using the analogy of medical testing [44], precision is the predictive value of a positive test (the proportion of positive tests that identify diseased persons); recall is analogous to the sensitivity of a test (the proportion of diseased persons a test classifies as positive). The term "specificity" is used in both information science and medicine. Specificity is the proportion of nonrelevant citations not retrieved (the proportion of nondiseased persons the test classifies as negative), or d/b+d.

In 1981, Fenichel [45] published a review of research related to the search process. Included are studies about professional searchers and end users, and MEDLINE

and other search systems. She provides the following useful table (Table 2) to summarize the quantitative information describing the "average" search based on the studies she reviewed:

Table 2

Summary Statistics About the "Average" Search

Variable	Range of Reported Means
Descriptors searched	3 to 47
Commands used	10 to 52
Connect time	6 to 37 minutes
Retrieved references	10 to 185
Recall	41 to 61 percent
Precision	17 to 81 percent
Unit cost	0.48 to 4.5 min/relevant reference

Source: Fenichel CH. The process of searching online bibliographic databases: a review of research. Library Research 1980-81;2:121.

In the same article, Fenichel describes her own doctoral dissertation [46] on the relationship between experience and both search process and outcome. Search process was measured by the use of system features (e.g., number of commands and descriptors used, connect time). Recall was one of the measures of search outcome. She found that novices did surprisingly well, although, as a group, they searched more slowly, made more errors, and had lower scores on the outcome variables compared to more experienced searchers. Those with the most experience scored high on search process and recall, "suggesting a relationship between search effort and recall." She found considerable variation in search behavior, even among people with the same experience.

In a series of articles titled "A Study of Information Seeking and Retrieving" [47], Saracevic and Kantor describe a large study of information retrieval. They analyzed five classes of variables (users, questions, searchers, searches, and items retrieved) to identify those that increased the odds of increasing relevance, precision, and recall. Pertinent to the proposed study are the following findings: relevance odds increased for questions users considered well-defined; when users had high expectations of finding information, relevance and precision increased. Judges had fair agreement in characterizing questions as to degree of clarity, specificity, complexity, and presuppositions. They detected no effect of frequency of searching, but, in their study, all of the subjects were experienced searchers. Overlap in choice of search terms by different searchers was low; the overlap of retrieved items was lower. The same items retrieved by different searchers for the same question were more likely to be relevant. The mean values for precision (57%) and recall (22%) were not inversely related, as anticipated, but slightly correlated. The authors question the utility of the measures of precision and recall. The analysis of the relationship between relevance of items retrieved and given variables was more powerful than the analysis of the relationship between precision and recall with given variables.

As part of a study to assess end-user searching, Poisson [48] analyzed the results of searches on a given topic completed by five end-user physicians. An experienced searcher defined the number of relevant citations in the database. Poisson found that precision was "quite good" but recall showed "more dramatic differences in the quality of these searches," with a range in recall of 12%-80%. Users with poor recall ratios tended not to use MeSH subject headings, subheadings, or explosions (a way of

including related terms), and they relied on textwords. She detected no relationship between performance and experience searching. The best searchers had previous instruction from librarians, who had emphasized MEDLINE system features.

Wildemuth and Moore [49] analyzed the transaction logs of 161 searches done by third-year medical students. Performance for 58 searches was determined by three factors: (1) 1ibrarians' ratings of four dimensions of quality of the search based on the transaction log (e.g., correct use of system syntax), (2) students' ratings of satisfaction with their searching behaviors, and (3) librarian identification of missed opportunities. The study did not find strong links between search behaviors and ratings of search performance. The most common missed opportunity was failure to take advantage of the controlled vocabulary, MeSH.

Haynes and his coworkers at McMaster University have contributed significantly to the study of MEDLINE and its use. They conducted a large study of MEDLINE use by 158 physicians and trainees at a teaching hospital [50]. Participants received three hours of instruction and free access to MEDLINE in the hospital setting. Researchers interviewed a random sample of the searchers about the perceived value of the searches and effect of search results on clinical decisions. Based on 280 interviews, searchers indicated that they were moderately satisfied with their performance and results of their searches (4.1 and 4.0 on a scale of 7, respectively). Participants indicated some improvement in patient care for 92% of the searches performed in response to a clinical problem (average score 3.1 on a scale of 1 to 7, with 1=not at all, 7=important improvement). Fifteen percent of the citations retrieved were rated as directly relevant to the question.

To determine searchers' competencies, the McMaster researchers randomly selected search questions, which were sent to one of three librarian search experts and to 1 of 13 clinicians with search experience. Both completed MEDLINE searches based on the questions. Citations that appeared in only one of the three searches (original, search expert, clinician searcher) conducted were evaluated for relevance on a 7-point scale by a clinician considered an expert in the subject area of the search. Relative recall (number of relevant citations from a search divided by total number of relevant citations from all three searches) and precision (number of relevant citations in the search divided by total number of citations in the search) were reported for each of the three groups [51]. Relative recall, rather than recall, was used since it is nearly impossible to determine the total number of relevant citations in a huge and continuously updated database. Seventy-eight searches were analyzed. Novice searchers had the lowest recall and precision. Librarians had the highest precision. Experienced clinician searchers and librarians had similar recall rates. As was found in the NLM Critical Incident Technique study, novice searchers depend more on text words and less on the more advanced features such as the MeSH vocabulary.

Surprisingly, only 4% of relevant citations were retrieved by all three searchers (novice, experienced clinician searcher, librarian). When three librarians performed five searches, there was an overlap of only 25% of the relevant citations. There were notable differences in searching style among the three librarians, based on variables such as number of terms used, and the number of times features such as MeSH, subheadings, and limit capabilities were used. Despite these differences in searching styles, there were no important differences in recall and precision for the three librarians.

Apparently, searchers can use very different strategies and retrieve different citations yet still have similar recall of relevant citations.

In a study to compare the performance and cost of 11 commercially available MEDLINE systems with MEDLINE from the National Library of Medicine, the McMaster group pooled results of searches by librarians and clinicians and rated the pooled results for relevance on a 7-point scale [52]. They found that the number of relevant and irrelevant citations retrieved by clinicians was higher, as was the total number of citations, compared to searches done by librarians. In a study to determine the impact of user fees [53], Haynes et al. used measures of precision and recall to compare searches done by those who had to pay versus those who searched for free. They found no difference in the quality of searches among the two groups.

The research conducted by Pao [54,55] and colleagues is directly related to the proposed study and will be discussed in detail. Their medical school's curriculum did not require students to have formal MEDLINE training but students did have convenient and free access to MEDLINE. The researchers asked: (1) Is there a relationship between experience searching (prior use) and performance on a search assignment? (2) Is there a relationship between searching experience and subsequent use of the database? (3) Would students acquire the skills and habit of searching without a formal course but with ubiquitous access to the system and a required assessment of their search skill?

The campus search system enabled the researchers to inventory students' previous search experiences. They categorized three levels of search experience: beginner, intermediate, and advanced. The definitions of these categories translate into "less than one search session per month," "one search session per month," and "more

than one search session per month," respectively, for the 30 months prior to the assignment. It was noted that the number of search sessions may or may not be the same as the number of searches.

Students searched for information about one of three written patient scenarios during a timed Comprehensive Clinical Assessment exam. The scenarios were based on actual cases, and the three topics varied in complexity. The authors do not state how long students were given to complete their searches, but each component of the exam was allotted between 15 and 30 minutes. The students were instructed to search MEDLINE and print several abstracts containing information relevant to making a decision or confirming a point in the patient scenario. A librarian, who performed a high-recall search, and two subject domain experts, who performed high-precision searches, had previously searched the topics. The results of those searches were pooled and, after a few weeks, the subject experts were asked to rate the citations as relevant, partially relevant, or not relevant. Pao makes the distinction between *normal relevance* and *strong relevance*:

> Items judged either relevant or partially relevant are the basis for recall and precision computations using *normal relevance*. *Strong relevance*, however, is applied when only items judged to be of definite relevance are used for comparison. The latter obviously represents a more rigorous requirement. Strong relevance is probably the more appropriate standard to compare an average search conducted for clinical purposes. [56]

Pao and her colleagues analyzed students' search skills and search effectiveness. Search skills were measured by the use of search features including MeSH headings, limits to title words, language, year of publication, review articles, abstracts, human studies, and subheadings. All of the students used MeSH headings; use of the other

features varied. No statistically significant difference was detected when the use of features was compared by level of search experience. Recall performance and the ability to retrieve relevant items measured search effectiveness. Recall increased with level of experience; precision also increased, but to a lesser degree. This was true when either normal relevance or strong relevance was the criterion. Eighty-three percent of the students retrieved at least one item of definite relevance. A strong statistically significant difference was found for the ability to retrieve definitely relevant items among the searchers with different levels of experience.

Also, 73.4% of the students attained a precision level of 25% or more when the criterion was normal relevance; 60.3% attained a precision level of 25% or more when the criterion was strong relevance. The authors' conclusion was that, even with little formal MEDLINE training, most students could retrieve relevant items. They found no correlation between searchers with recall greater than 25% and either National Board Part II scores or scores on the Comprehensive Clinical Assessment exam. They did detect strong positive correlation between more search experience before the exam and the number of online sessions afterwards.

Shelstad and Clevenger [57] studied the ability of third-year medical students to answer clinical questions using the NLM's Grateful Med end-user search software. In response to two surgical questions, students submitted their strategies and retrieval, which were compared with "gold standard" searches run by an experienced medical librarian. For both questions, the majority of students failed to retrieve the number of citations in the gold standard range. Analyses of search strategies indicated that students did not retrieve relevant citations because they "used incorrect MeSH headings, failed to

explode MeSH headings, used inappropriate subheadings, used textwords instead of MeSH headings, or limited their searches to include only review articles" [58]. The authors conclude that there is a need for better models of instruction in information retrieval.

In a recent study, Burrows and Tylman [59] evaluated third-year medical students' search skills using a combination of search strategy and retrieval criteria. Three librarians completed a search on a clinical question and, with a fourth librarian, determined the important elements of an appropriate search strategy. This librarian also reviewed citations and abstracts retrieved by librarians and medical students and selected nine articles that contained information needed to make the best clinical decision.

Students' searches were ranked excellent, good, fair, and poor based on these criteria. Only 26% of the students' searches were ranked either good or excellent using the retrieval criteria. None of the search strategies was ranked excellent, and only five percent were considered good.

In summary, a review of the literature reveals several studies about MEDLINE and its use. Many researchers examine transaction logs to assess search skill. Search effectiveness is usually based on an analysis of retrieval. Relative recall is a common measure of retrieval. Topical relevance (relatedness) appears to be more often used, although some studies rely on self-reported user satisfaction to evaluate relevance. At least one study employs the criterion of strong relevance. The test search questions researchers used have been either spontaneous clinical questions or based on spontaneous clinical questions. There is some experience assessing medical students'

search performances under timed conditions. Repeated conclusions are that: even novices can retrieve relevant items; experience improves search performance; and failed searches can be attributed to the lack of use of search features, especially MEDLINE's structured vocabulary—MeSH.

Assessment in Medical Education

Any method for evaluating a medical student's search effectiveness must be accomplished within the context of the medical school curriculum. Its feasibility for implementation will depend on its congruence with other assessment activities and the prevailing culture of medical education. Currently, there is nationwide discussion about medical school curricula and widespread agreement that change is necessary. It will be useful to review, briefly, some of the major trends in medical curriculum reform since they influenced the design of this study.

There is general agreement that assessment in medical education is weighted too heavily toward a student's ability to recall facts rather than demonstrate skills. Those advocating reform in medical education want to balance knowledge acquisition with increased emphasis on the development of skills, behaviors, and attitudes that encourage lifelong learning. Examples of skills considered important include clinical competence appropriate for an undergraduate medical student; problem-solving skills; and the ability to acquire, evaluate, and apply new information. The all-too-common multiple-choice test may measure knowledge acquisition, but does not test skill development. Methods of assessment should be consistent with what is being assessed; reform in education should be matched with reform in evaluation.

The results of a 1989 survey of 1,369 faculty and administrators from U.S. medical schools indicate that 96% of respondents support testing mechanisms to evaluate a student's independent problem-solving skills. Ninety-seven percent support increasing the integration of the basic sciences and clinical phases of medical education; 86% support decreasing the number of large lectures and increasing student time for independent student and interaction with faculty [60].

Problem-based learning (PBL) is an increasingly popular method of educating medical students. The appeal of PBL is that it fosters many of the changes that medical educators advocate. Students in PBL programs actively participate in small groups to solve clinical problems that require the demonstration of an understanding of basic science concepts [61]. But, even when a medical school converts its entire curriculum to a PBL format, students still must pass the national licensing exam, a standardized test based primarily on the recall of facts. A targeted review [62] of twenty years of literature on the effectiveness of PBL suggests that students in PBL programs perform as well, and sometimes better, on clinical examinations and faculty evaluations. There is also some evidence that PBL students may do less well on basic science exams and that they feel less prepared in the basic sciences than do the non-PBL students. No one has been able to determine a significant advantage of PBL versus non-PBL curricula using standard measures of knowledge acquisition, and measures specifically designed to assess the medical problem-solving process have not been developed [63].

"Evaluation drives the system" [64]. The United States Medical Licensing Examination (USMLE) is the exam that determines whether or not a physician can practice in this country. There are three parts to the exam: Step 1 covers the basic

sciences and is taken by second-year medical students; Step 2 covers core clinical sciences and is taken by fourth-year students; Step 3 covers clinical practice and is taken by postgraduate resident physicians. Recently, the Federation of State Medical Boards of the U.S. and the National Board of Medical Examiners, joint sponsors of the program, made major changes in the USMLE. Beginning in 1999, the test was administered as a computerized exam, and students scheduled time to take it at commercial test centers. This format is expected to increase security and scheduling flexibility, and allow for enhancements in the testing method. Step 2 includes more multiple-choice patient care scenarios. Eventually, the USMLE will be an adaptive test in which the test content varies in response to the evidence of the knowledge of the student taking it.

Another trend in assessment of medical students is the growing use of the Objective Structured Clinical Exam (OSCE). OSCEs purport to assess clinical performance (skills) rather than knowledge acquisition (recall) [65]. The premise is that the "closer an examination simulates the eventual task, the more valuable the assessment will be" [66]. Typically, an OSCE consists of multiple "stations" where students must complete an activity during a specified time period. Examples of activities at OSCE stations include: perform a limited examination on a simulated patient, interpret an EKG, and write lab orders—activities practicing physicians routinely perform. OSCE exams take considerable time to develop and administer. OSCEs usually last several hours and require several stations to achieve acceptable levels of reliability.

Some of the studies discussed earlier describe MEDLINE assessment in the context of OSCEs. The Comprehensive Clinical Assessment Exam described in Pao's study is an objective structured clinical exam. Burrows and Tylman' developed their search question for an OSCE. One other publication reports assessing MEDLINE skills in the context of an OSCE, but not in a research setting [67]. In 1997, this researcher presented data on experience with a search station in an internal medicine clerkship OSCE [68].

The OSCE has many of the features of the proposed method for assessing a student's MEDLINE search effectiveness. OSCEs are typically criterion-referenced timed tests that are presented in a clinical context. Numerous papers describe the development and use of OSCEs, including methods to assess a test's validity and reliability, and ways to determine cutoff scores or passing grades [69-71]. Most of the methodologies described in those studies come from the more general literature of educational assessment, which was consulted for this study.

In summary, trends in medical student assessment support the use of tests that evaluate skills or performance in a clinical context. A student's skill in using computers is becoming increasingly more important. The objective structured clinical exam is a good model for the development of a method to assess MEDLINE search effectiveness.

CHAPTER 3

METHODOLOGY

Purpose of the Study

The purpose of this study was to develop and evaluate a method for assessing the ability of medical students to retrieve information relevant to patient care questions using the MEDLINE database. Specific objectives were to:

1. determine if this method of assessment can identify students with varying levels of search effectiveness and students with unacceptable levels of search effectiveness;

2. recommend the number of searches needed for an acceptable level of reliability;

3. test reliability and validity across two popular search systems (Ovid and OvidWeb);

4. examine the relationship between test score and level of education, previous MEDLNE training, and previous MEDLINE search experience.

Design

This study is an evaluation of a proposed method, which was designed as the MEDLINE Metric, for assessing medical students' search skills. MEDLINE Metric consists of: (a) the development, by experts, of realistic clinical scenarios that include highly constructed search questions designed to test defined search skills; (b) timed tasks (searches) completed by subjects; (c) the evaluation of search results; and (d) instructive feedback. A goal is to offer medical educators a valid, reliable, and feasible way to judge

mastery of information searching skill, which is one of the informatics skills, by measuring results (search retrieval) rather than process (search behavior) or cognition (knowledge about searching). The assessment is designed to determine mastery in MEDLINE searching. It is a criterion-referenced, rather than a norm-referenced, evaluation (i.e., scores are based on mastery of a task as defined by experts rather than on performance of typical people) [72].

In their book, *Evaluation Methods in Medical Informatics*, Friedman and Wyatt [73] distinguish between measurement and demonstration studies and note the importance of measurement studies to "determine with how much error an attribute of interest can be measured in a population of objects . . . Measurement procedures developed and validated through measurement studies provide researchers with what they need to conduct demonstration studies [74]. In this measurement study, the attribute of interest is the search effectiveness of medical students as measured by their ability to retrieve specific articles from the MEDLINE database that are strongly relevant to carefully constructed clinical scenarios with search questions. The target population is medical students. Clinical scenarios, rather than isolated search questions, are used for two reasons. Scenarios provide context for the question, and thus help specify the relevance of retrieved citations. The other reason is that medical licensure exams include an increasing number of clinical scenarios and medical students are familiar with this format.

An overview of the MEDLINE Metric development process is illustrated in Figure 2.

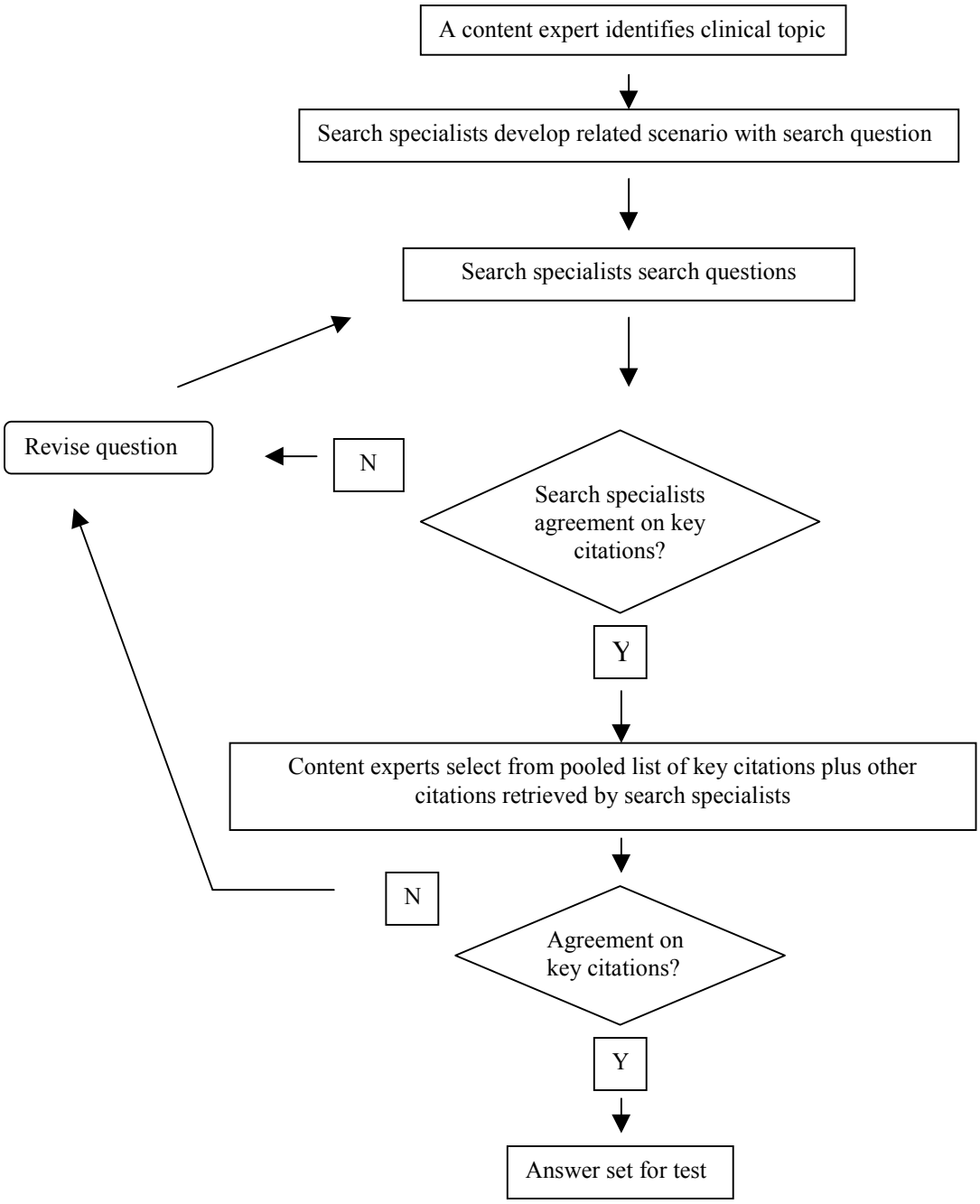

Figure 2. MEDLINE Metric development process.

Scenarios, questions, and key citations were determined in consultation with four medical content experts and four MEDLINE search specialists. The medical content experts were medical school faculty members with teaching responsibilities. They

included a pediatrician, a family medicine practitioner, and two internists, with an approximate total of 50 combined years of clinical experience. The MEDLINE search specialists were medical librarians from three different medical schools and one affiliated teaching hospital. These librarians teach MEDLINE searching and have a combined total of more than 50 years of experience searching the database. The author coordinated the process of developing scenarios and search questions, and identifying key citations.

Three of the four medical content experts provided suggestions for realistic clinical scenarios that would prompt searches for information found in medical journal articles. They were asked to consider "core topics," such as those outlined in the *Core Medicine Clerkship Curriculum Guide* [75], common diseases and conditions that medical students are expected to be familiar with from their third-year clerkships. Content experts were also given examples of scenarios that have been used at Texas A&M College of Medicine to assess MEDLINE skills in the third-year Internal Medicine Clerkship Objective Structured Clinical Exam (OSCE) (see Appendix A).

Three of the four MEDLINE search specialists, with this researcher, identified the types of search skills they thought would constitute basic proficiency in searching MEDLINE (see Appendix B). These skills include the ability to: (a) use the MeSH vocabulary, (b) limit retrieval, (c) use subheadings, (d) combine concepts, (e) specify terms by database fields, and (d) select articles relevant to a target audience (Table 3). Articles cited in the literature review, as well as books and syllabi for teaching searching, were consulted to identify core search skills [76-78].

Table 3

MEDLINE Core Search Skills

Use controlled vocabulary (Medical Subject Headings—MeSH)
Limit retrieval (e.g., to human subjects, by language, by publication type/year)
Use subheadings
Combine concepts
Specify by database field (e.g., author, title word)
Select articles most relevant to target audience for information

Using the topics identified by the medical content experts, these search specialists developed short clinical scenarios and search questions designed to necessitate the use of one or more of the defined basic search skills in order to retrieve citations with strong relevance efficiently. Strong relevance, according to Pao [79], describes items "judged to be of definite relevance" and "is probably the more appropriate standard to compare an average search conducted for clinical purposes." For example, a question might ask for treatment information specific to a certain age group. To complete the search successfully and efficiently, the student will need to know how to limit search results by age group, a feature of MEDLINE systems. Based on the topics identified by the medical content experts, each search specialist was asked to write scenarios and questions.

Eleven scenarios with questions were developed. These scenarios and questions were then distributed to the search specialists who searched MEDLINE and identified three to five most relevant citations for each question. The search specialists critiqued the scenarios and questions and commented on the search skills required to retrieve relevant citations. This researcher reviewed search results and, in consultation with the others, modified the questions. As illustrated in Figure 2 (MEDLINE Metric Development

Process), modification of the questions continued until all of the search specialists had some citations in common. At this point, a fourth expert MEDLINE search specialist, new to the study, was recruited to complete the searches and provide a source of external validity since the original three search specialists had searched variations of the same questions several times. With some minor modification to the scenarios and questions, six searches completed by four search specialists resulted in retrieval that included common citations for each. A matrix was created to identify which search skills were addressed by each question (see Table 4).

Table 4

Search Skills Expected to Result in Efficient Search and Relevant Retrieval for Each Test Question

QUESTION	SKILLS					
	Use MeSH Headings	Limit Retrieval	Use Subheadings[1]	Combine Concepts	Specify Field	Select for Audience
Q1 Calan	Verapamil Neoplasms	Human English	CI, AE, ET	Yes	No	Patient
Q2 CT[2]	CT Appendicitis	English	DI, EC	Yes	No	Administrator
Q3 NSAIDS[2]	GI Hemorrhage NSAIDs	English Age Group	CI,ET,AE	Yes	No	Resident
Q4 Diabetes	Diabetes Mellitus United States	English	EC	No	No	Faculty
Q5 Otitis	Otitis Media Amoxicillin	Human English	PC	Yes	No	Pediatrician
Q6 Ingelfinger	No	Year	No	No	Author Title Keyword	Student

[1]Subheadings: CI (chemically induced), AE (adverse effects), ET (etiology), DI (diagnosis), EC (economics), PC (prevention and control)
[2]CT (computerized tomography)
[3]NSAIDS (non-steroidal anti-inflammatory drugs)

The goal was to develop scenarios and questions that require basic search skills and result in the retrieval of citations that have strong relevance to the question posed in the scenario. They will also be citations that multiple search specialists have retrieved using different search strategies. That is, they also will be highly retrievable. These highly retrievable, highly relevant citations are called "key citations" in this study.

For each question, the citations identified as most relevant by each search specialist were combined into one list. The key citations (those that all search specialists retrieved) were not distinguished in the combined list. To further verify the relevance of the key citations, two medical content experts reviewed these lists, including a fourth content expert who had not been previously involved in the study. They were asked to identify a specific number of most relevant citations for five scenarios and questions. The sixth scenario and question were for "known articles"; that is, the key citations were the only (and most relevant) citations retrieved by all of the search experts. The specified number to select was either two or three; four of the scenarios had one key citation all search specialists retrieved, one scenario had two key citations. The number to select was based on the number of key citations retrieved by all search specialists and the number of other relevant citations retrieved by the individual search specialists. The selections of the medical content experts included the key citations for each of the five scenarios, thus reinforcing confidence that unanimous agreement among the search specialists had resulted in citations that were strongly relevant from a clinician's perspective as well.

In summary, determination of key citations was a two-step process. Questions were revised until search specialists agreed on key citations. Medical content experts then selected from a combined list of searchers' citations, including the key citations. If there

had not been unanimous agreement about the key citations, the scenario and question would have been revised until there was agreement or the question would have been eliminated altogether. The resulting set of clinical scenarios, search questions, and key citations were the basis for a test to assess students' search effectiveness. The scenarios and questions made up the test bank; the key citations were the "answers." A fourth-year medical student took the test and verified that the questions were clearly stated and not too difficult for a student to search. During the question development phase of the study, two fourth-year medical students demonstrated the feasibility and logistics of test administration (timing, selection, and printing) by completing searches on sample questions within a specified time limit.

The MEDLINE Metric development process resulted in a six-item test. In phase one of this study, the test was administered to 113 people with different levels of education, MEDLINE search experience, and training. Ten variations in the order of the six searches were used to minimize possible bias related to the learning effect of any particular search and to any fatigue factor in the overall test. In each scenario, students were instructed to identify a specified number of citations they considered most relevant to the question. Two questions asked for two citations each; four questions asked for three citations. A sample question is given in Figure 3.

You are a pediatrician and see lots of middle ear infections. For patients with recurrent infections, you wonder about the effectiveness of using antibiotics, specifically amoxicillin, for prevention. You worry about the overuse of antibiotics; is it better not to use them to prevent these infections? Find articles that provide relevant data. Use the most current file of MEDLINE (1995-1998). Select two articles that are most relevant.

Figure 3. Sample question.

Based on the experience with phase one of the study, changes were made in the test for phase two. Subjects were instructed to select two to five most relevant citations, to use "the most current file of MEDLINE," and to select articles published before 1999. Figure 4 illustrates the changes made in the instructions.

You are a pediatrician and see many middle ear infections. For patients with recurrent infections, you wonder about the effectiveness of using antibiotics, specifically amoxicillin, for prevention. You worry about the over use of antibiotics; is it better not to use them to prevent these infections? Find articles that provide relevant data. Use the most current file of MEDLINE. Select 2-5 articles that seem most relevant and that were published before1999.

Figure 4. Sample revised question.

Previous research, personal experience, and consultation with search specialists suggested that 20 minutes is a reasonable amount of time to allot for each search. Subjects who completed a search in less than 20 minutes were allowed to start a new search to be completed within 20 minutes; the subject was not allowed to use time "saved" from one search to extend the length of time to complete another search.

In addition to performing six searches, each subject was asked to complete a Human Subjects Consent Form, a pretest questionnaire, and six posttest questionnaires. A sample test packet for the first phase of the study can be found in Appendix C. Student

subjects received written feedback, including test score and hints for improving search effectiveness (see Appendix D). Figure 5 provides an overview of the test administration procedure.

The test administration procedure was the same in phase one and phase two.

Population and Test Group

The target population for the proposed method of assessment is fourth-year (senior) medical students in the United States. Students at this level will have experience with clinical questions, most likely some MEDLINE search experience, and perhaps some formal MEDLINE training. By this point in their medical education, MEDLINE searching is no longer an academic exercise, but a useful skill for medical practice. Senior medical students are expected to be able to locate literature relevant to patient care. A valid and reliable method for assessing MEDLINE search effectiveness would be appropriate in the fourth-year curriculum.

The selection of subjects to test the MEDLINE Metric test instrument was purposive, designed to include people with a wide range of MEDLINE search skills. In phase one of the study, subjects included 7 undergraduate biomedical sciences students, 64 second-year medical students (M2s), 2 third-year medical students (M3s), 29 fourth-year medical students (M4s), and 11 medical librarians (MLSs) who are experienced searchers. The second-year medical students completed the study as part of a required preceptorship course activity. All other students were recruited to complete the study and were paid $15 each for their participation. Flyers were used to advertise the study to these

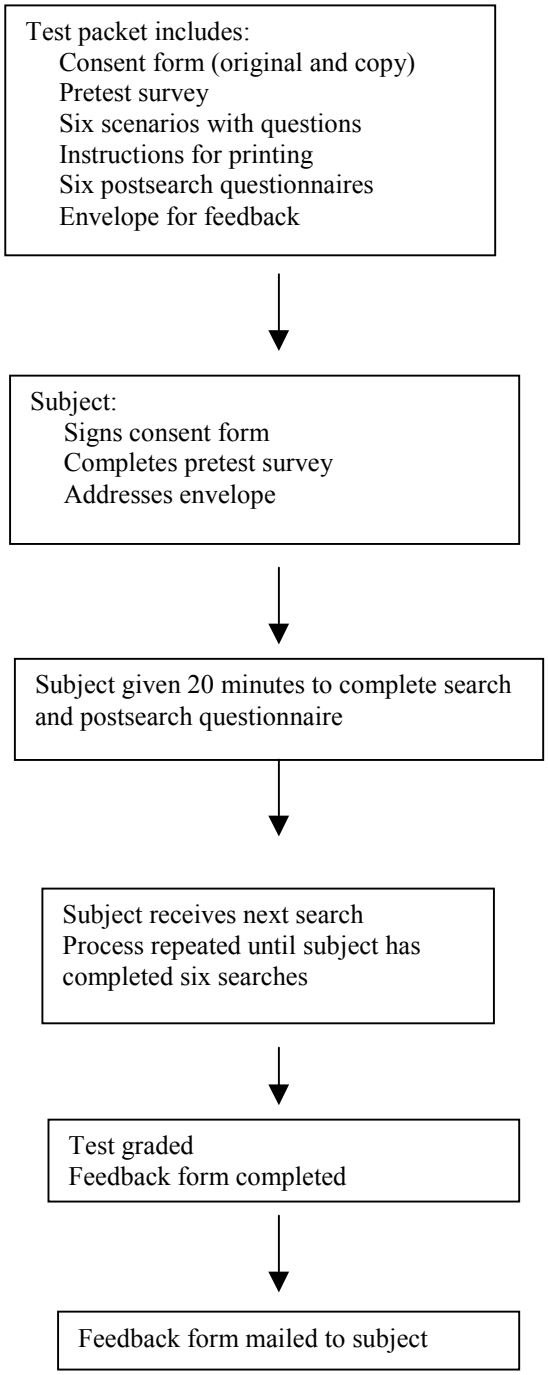

Figure 5. Test administration procedure.

.

students. Computers and Medical Information, an elective taken by fourth-year medical students, was the primary source of student subjects. Librarian participants were selectively recruited, based primarily on referral by known search experts. An electronic timer was sent to those librarians who agreed to participate, as a tool for taking the test and a token of appreciation for their time. Phase two of the study added 22 fourth-year students and 10 medical librarians. In phase one as well as phase two, all subjects signed and received a copy of an approved Informed Consent Form.

Data Collection

Each subject completed a Pre-Search Questionnaire, which provided information about grade-level, formal MEDLINE training, previous MEDLINE search experience, and the search system used for the test. Subjects were instructed to complete each of six searches within a 20-minute time period. In the class environment, subjects had all of the questions in a packet. They were told when the 20-minute time periods were up and instructed to take out the next search. When the test was administered to an individual or small number of subjects, the questions were distributed one at a time as subjects completed their search or the 20-minute time period elapsed. Librarian subjects were instructed not to look at the scenarios beforehand, and to keep track of time on their own.

For each search, subjects were instructed to identify a specified number (phase one) or range (phase two) of the most relevant citations they found. They were also asked to rate (on a scale of 1 to 5) their satisfaction with their results, to indicate if the search was difficult, and to make comments about the search or the activity.

Supervised testing of students took place in a medical library or learning resources center. Volunteers made appointments with a designated library or learning

resources center staff member to take the test. These staff members distributed the Informed Consent Form, the Pre-Search Questionnaire, each of the test questions, and the Post-Search Questionnaire. They monitored test time and collected search results and Post-Search Questionnaires at 20-minute intervals. The tests were graded using the key citations as the answer set. Coding forms were completed for each subject, including: (a) study identification number, (b) information about level of education, (c) formal MEDLINE training, (d) number of MEDLINE searches done in the past six months, (e) MEDLINE system used during the test, and (f) test variation number. For each search, satisfaction with results and perceived difficulty of the search, comments, and grade were recorded. See Appendix E for the coding form.

Data Analysis

The score for each search is the number of key citations retrieved divided by the total number of key citations for that search, or the proportion of key citations in the retrieved set. An individual student's total score is the average of the six search scores. During the course of the study, one of the key citations for Question 3, published in 1995, was dropped from the current file of MEDLINE. To be fair and consistent in scoring throughout the testing period, all subjects received full credit for identifying either of the two original key citations.

Frequency distributions were plotted for total scores and for the scores of each of the six searches. Descriptive statistics (means, medians, standard deviations) and standardized scores were calculated.

For each research question the following analysis was made:

1. Is the proposed method for assessing search effectiveness valid?

The determination of content validity for the test is of primary importance. Content validity is "primarily a matter of professional judgement on the part of teachers and subject-matter specialists" [80]. Search specialists defined the instructional objectives assessed by the test using their own experience teaching medical students plus information from books and syllabi intended to teach end users how to search MEDLINE. Content experts created realistic scenarios that would prompt a MEDLINE search.

2. Does the test discriminate among different performance levels?

A valid method of assessment should distinguish among presumed levels of expertise. Analysis of variance was used to determine if scores differed based on three characteristics: level of education, search training, and previous search experience. One would expect higher scores for those with more education, formal training, and search experience.

Logistic regression was used to determine the degree to which test scores appropriately categorize subjects by level of education, search training, and search experience. Analysis of test scores provided an estimate for a cutoff score to define acceptable mastery of MEDLINE search skills.

3. Are the results of the test reliable?

Internal consistency indicates how well different items measure the same attribute [81]. Cronbach's coefficient alpha was calculated to assess internal consistency of the six-item test. That formula is [82]:

$$\alpha = (n/n-1)[\ (SD_t^2 - \Sigma SD_i^2) / SD_t^2\ ]$$

where α is the estimate of reliability,

n is the number of items in the test,

SD_t is the standard deviation of the test scores, and

SD_i is the standard deviation of the scores from a group of individuals on an item.

Expressed another way [83], the formula for Cronbach's alpha is:

$$\alpha = (n/n\text{-}1)x \frac{\text{variance of total scale—sum of variances of individual items)}}{\text{variance of total scale}}$$

The reliability of an entire test can also be estimated from an analysis of the statistics of the individual items, as if each item constituted a parallel test. Estimating reliability depends on the consistency of the performance of an individual and is based on the standard deviation of the test and the standard deviations of the items. Reliability was also to be examined by comparing scores of students who were tested and then retested.

4. How many test searches are needed to reliably assess a student's performance level?

The reliability of a test depends on its length. The relationship between reliability and test length was calculated using the Spearman-Brown Prophecy formula [84].

$$r_{kk} = \frac{kr_{tt}}{1 + (k - 1)\, r_{tt}}$$

where r_{kk} is the reliability of the test k times as long as the original test,

r_{tt} is the reliability of the original test, and

k is the factor by which the test length is changed.

Expressed another way, the Spearman-Brown Prophecy formula is [85]:

$$\frac{\text{k(average correlation among all items)}}{1 + (\text{k} - 1)\text{ average correlation among all items)}}$$

Feedback about the administration of the test during the study provided information about the feasibility of administering the test as part of a medical school curriculum and the possible impact of changing the length of the test.

5. Is the method of assessment reliable across different MEDLINE search systems?

Analysis by t-tests determines any difference in average total scores based on the search system that was used. Calculating Cronbach's alpha for each subset (system) provides a measure for comparing reliability.

To answer the next questions, correlation coefficients were calculated.

6. What is the relationship between level of education and test score?

7. What is the relationship between previous MEDLINE search training and test score?

8. What is the relationship between previous MEDLINE search experience and test score?

The relationship between satisfaction with search results and test score was also analyzed.

CHAPTER 4

FINDINGS

Test Subjects

One hundred forty-five subjects completed the test between August 1998 and January 2000. Data from undergraduates and third-year medical students were eliminated from analysis because of the small sample size in each category. No resident physicians volunteered to take the test despite several efforts to recruit them.

Only complete cases were used. If one or more of the six search questions' results was missing, the entire case was eliminated from analysis. Missing data typically occurred when students did not turn in search results due to either printing or e-mail problems. Table 5 categorizes the subjects and valid cases by level of education.

Table 5

Number Tested and Valid Cases

Test Subjects	Tested Frequency	(%)	Valid Cases Frequency	(%)
Undergraduates	7	(4.8)	-	-
M2 Second-year medical students	64	(44.1)	53	(44.2)
M3 Third-year medical students	2	(1.4)	-	-
M4 Fourth-year medical students	51	(35.2)	47	(39.2)
MLS Medical librarian searchers	21	(14.5)	20	(16.7)
TOTAL	145	(100.0)	120	(100.0)

As expected, the reported amount of formal MEDLINE training highly correlates with each subject's level of education (Spearman *rho* .754, $p<.001$, 2-tailed) Most second-year medical students (72.5%) reported less than two hours of MEDLINE.

training; 15.7% reported no formal MEDLINE training, which is contradictory to fact since a two-hour MEDLINE session was required during their first-year medical biochemistry course. The rest (11.8%) indicated they had two to four hours of training. Fourth-year medical students reported training that was more varied: 4.3% indicated no formal MEDLINE training; 25.5% indicated less than two hours; 40.4% indicated two to four hours; and 29.8% reported more than four hours of training. By their fourth year, students have had several opportunities for MEDLINE training in the curriculum. All of the medical librarians (100%) indicated that they had more than four hours of MEDLINE training.

Before taking the test, each subject recorded the approximate number of MEDLINE searches done in the past six months. The numbers vary greatly (from 0 to 900). Since the ranges are wide, it is useful to look at the median number, as well as mean number, of searches by educational level. These statistics are displayed in Table 6.

Table 6

Reported Number of MEDLINE Searches Completed During the Past Six Months

Education	Number of Searches				
	Mean	SD[4]	Minimum	Maximum	Median
M2[1]	2.971	4.827	0	25	1
M4[2]	10.202	13.829	0	88	6
MLS[3]	227.444	234.454	20	900	127.5
TOTAL	40.733	121.171	0	900	5

[1]M2 - Second-year medical students
[2]M4 – Fourth-year medical students
[3]MLS – Medical librarian searchers
[4]SD- Standard Deviation

Search System

Subjects were asked to indicate the search system used for the test. Ninety-seven subjects (80.8%) used the Ovid Web search interface, nine (7.5%) used NLM's PubMed, and the rest (11.7%) used other MEDLINE search interfaces such as Ovid CD and Ovid Telnet. One person reported using more than one search system to complete the test.

Test Variation

The test consisted of six search scenarios and questions presented in one of ten variations. The difference in the variations was simply in the order of the questions. The distribution of test variations used in the study is illustrated in Figure 6.

Scoring of Questions

Individual question scores are the proportions of key citations retrieved for each of six search questions.

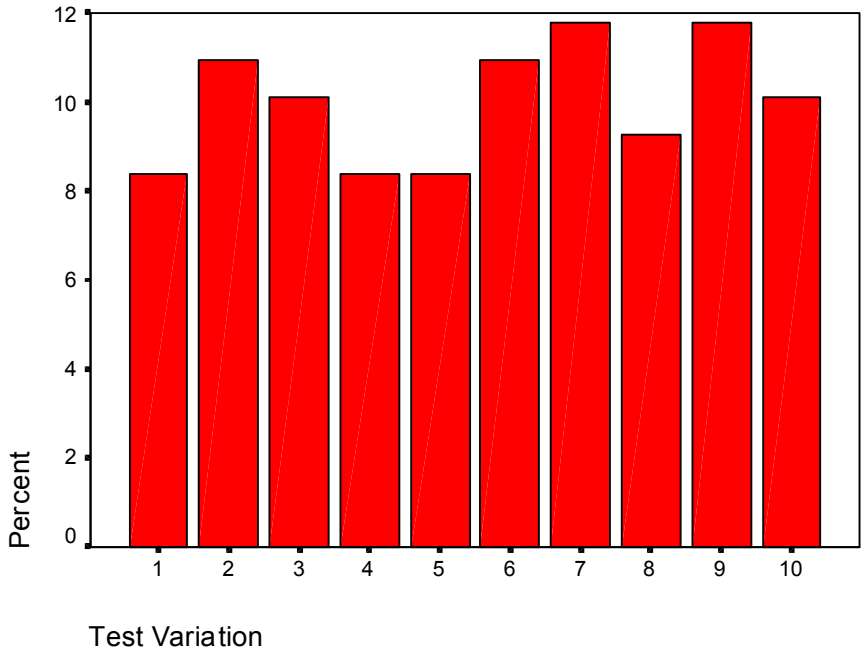

Figure 6. Distribution of test variations used in the study.

For five of the six questions, the frequency distribution of the scores was binomial. Table 7 shows the range, mean, and standard deviation of the scores for the six questions.

Table 7

Scores for Correctly Identified Citations

		Scores			
Question Number	Number of Subjects	Minimum	Maximum	Mean	SD
Q1	120	.00	1.00	.5917	.4936
Q2	120	.00	1.00	.6500	.4790
Q3	120	.00	1.00	.5000	.5021
Q4	120	.00	1.00	.6167	.4882
Q5	120	.00	1.00	.4000	.4920
Q6	120	.00	1.00	.5208	.4463

The difficulty index is the percentage of people who answered a question correctly of those who attempted to answer the question [86]. A lower percentage indicates a more difficult question. The difficulty index for the first five questions is the same as the mean; 41.7% of subjects received full credit—identified both key citations—for Question 6. Question 5 is the most difficult; Question 2 is the least difficult according to this criterion. Table 8 orders the questions according to difficulty, from most difficult to least difficult. An indicator of the question topic is provided.

Table 8

Questions Ranked by Difficulty

Rank	Question Number and Topic	Difficulty Index
1	Q5 (Otitis)	.40
2	Q3 (NSAIDS)[1]	.50
3	Q6 (Ingelfinger)	.52
4	Q1 (Calan)	.59
5	Q4 (Diabetes)	.62
6	Q2 (CT)[2]	.65

[1]NSAIDS - Non-steroidal anti-inflammatory drugs
[2]CT- Computerized tomography

Using a five-point scale (1-Strongly Disagree to 5-Strongly Agree), subjects rated their results by indicating agreement with the following statement: "I found highly relevant articles that I expect would answer the question asked." They also answered "Yes" or "No" to the question: "Did you have difficulty with this search?" Additionally, there was a place for comments on the post-search questionnaire. The following analyses examine these responses for each question.

Question 1: You are a patient with chronic heart disease who regularly takes Calan. A co-worker tells you he has heard that some of the drugs used for heart disease increase the risk of getting cancer. Should you be concerned? Find studies that address this. Use the most current file of MEDLINE. Select *2-5* articles that seem most relevant and that were published *before*1999.

Among all subjects responding to this question (n=119), 47.9% indicated some difficulty with the search. Overall, 74.8% of the subjects agreed (level 4 or 5) that they had found highly relevant articles to answer this question. The percent of subjects who felt they were successful increased with level of education (63.5% of M2s, 78.7% of M4s, and 95.0% of MLSs). Indication of finding relevant articles and score for this item were positively correlated (Spearman rho=.324, p<.001, 2-tailed).

Looking at the comments, more people reported problems finding enough articles than finding too many. A common comment had to do with figuring out that Calan, a term not in the MeSH vocabulary, was the same drug as verapamil, a MeSH term.

Question 2: You are an ER physician, and next week you meet with hospital administrators to propose the use of CT scans for patients suspected of having

appendicitis. You think that this diagnostic test would reduce the number of unnecessary appendectomies and reduce hospital costs. You want to find articles to cite to support your proposal. Use the most current file of MEDLINE. Select *2-5* articles that seem most relevant and that were published *before*1999.

Only 12.0% reported difficulty with this search (n=117). Overall, 90.7% of the subjects agreed (level 4 or 5) that they had found highly relevant articles to answer this question. By educational level, 78.9% of M1s and 100% of both M4s and MLSs agreed. Question 2 was the least difficult question, according to the difficulty index, although there was not a statistically significant correlation between indication of finding relevant articles and score for this question (Spearman *rho*=.152, $p<.101$, 2-tailed).

Selecting the right terminology was not a problem; according to their comments, more people had trouble finding enough articles than limiting their retrieval.

Question 3: You are a busy first-year pediatrics resident and just admitted a child with a GI bleed. You wonder if it could have been caused by her use of NSAIDs (non-steroidal anti-inflammatory agents). Find current articles that deal specifically with children under twelve. You don't have time to read case reports. Use the most current file of MEDLINE. Select *2-5* articles that seem most relevant and that were published *before*1999.

Of 119 subjects, 52.9% indicated some difficulty with this search. Slightly more than half (55.9%) of the subjects agreed that they had found highly relevant articles (level 4 or 5) for this question. By level of education, 56.8% of M2s, 44.6% of M4s, and 80.0% of MLSs agreed. Correlation between indication of retrieving relevant articles

and score for this question was statistically significant (Spearman *rho*=.205, *p*<.026, 2-tailed).

Not knowing how to limit the search to the pediatric population was a very common comment.

Question 4: Diabetes mellitus is a significant public health problem in the U.S. You have been invited to give a lecture to medical students about diabetes and you want to start with current statistics about the direct and indirect medical costs associated with this disease. Use the most current file of MEDLINE. Select *2-5* articles that seem most relevant and that were published *before*1999.

Only 13.6% of the subjects (n=118) indicated any difficulty with this search. Most subjects (88.1%) agreed (level 4 or 5) that they had found highly relevant articles to answer this question. There was not much difference based on educational level. In fact, M2s reported slightly higher agreement than M4s (86.2% vs. 85.1%, respectively). All of the MLSs agreed that their retrieval included highly relevant articles. There was not a statistically significant correlation between indicating agreement about finding relevant articles and the score for this question (Spearman *rho*=.135, *p*<.145, 2-tailed).

Several comments had to do with narrowing retrieval and feeling confident that they had selected relevant articles. Many articles did not have abstracts, which seemed to make selection more difficult.

Question 5: You are a pediatrician and see lots of middle ear infections. For patients with recurrent infections, you wonder about the effectiveness of using antibiotics, specifically Amoxicillin, for prevention. You worry about the over use of antibiotics; is it better not to use them to prevent these infections? Find articles that

provide relevant data. Use the most current file of MEDLINE. Select *2-5* articles that seem most relevant and that were published *before*1999.

Nearly 20% (19.2%) reported some difficulty with this search (n=120). Overall, 81.7% of subjects agreed (at level 4 or 5) that they had retrieved highly relevant citations. There was close agreement on this across education levels (77.4% of M2s, 85.1% of M4s, and 85.0% of MLSs). There was no statistically significant correlation between indicating satisfaction with the retrieval and the score for this question (Spearman *rho*=.098, $p<.285$, 2-tailed).

This was the most difficult search based on the difficulty index criterion. More subjects reported problems narrowing their search than finding enough relevant articles. A few people commented that this was a "complex," and "multifaceted" search question, and that it was difficult to insure that articles addressed all of the elements.

Question 6: Your attending mentions the "Ingelfinger policy," which you have never heard of. She says it's about the release of prepublication information to the media and that, back in the 1980s, Arnold Relman, editor of the *New England Journal of Medicine* wrote about it. She asks you to find the original articles. Select *2-5* articles that seem most relevant.

Half (50.4%) of the subjects reported difficulty with this search (n=118). Overall, 65.8% of subjects agreed that they had retrieved highly relevant articles to answer this question. There were noticeable differences by educational level (49.0% of M2s, 71.8% of M3s, and 95.0% of MLSs). There was a statistically significant correlation between impression of finding relevant articles and score (Spearman *rho*=.634, $p<.001$, 2-tailed).

Comments suggested that many people did not know how to search by author or that they needed to change databases (in the Ovid Web system) to get articles from the 1980s.

Total Scores

The total test score is the average of the six individual scores. Total scores ranged from 0 to 1 with a frequency distribution as shown in Figure 7.

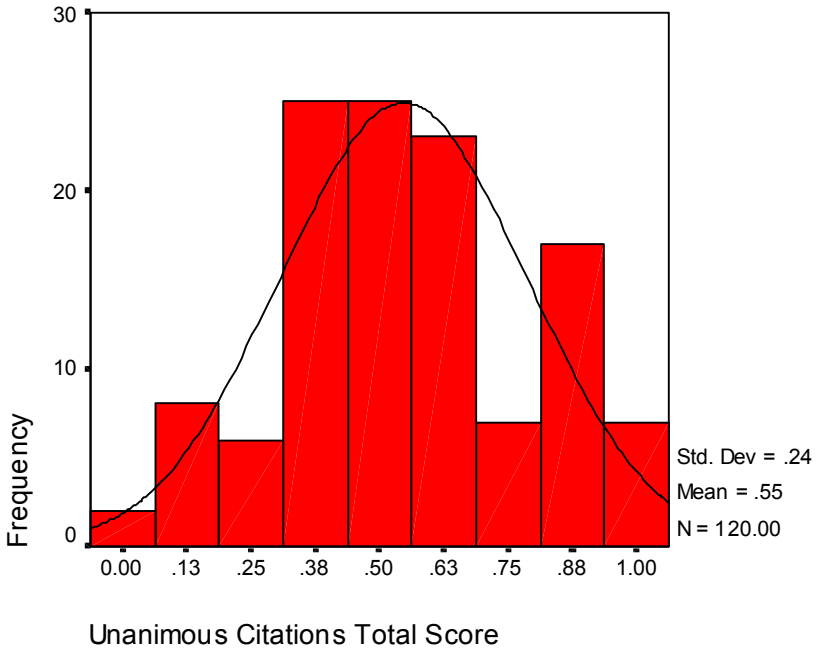

Figure 7. Distribution of total scores.

Total Score and Level of Education, Formal MEDLINE Training and Search Experience of the Test Subjects

In the pre-test questionnaire, subjects indicated level of education (second-year medical student, fourth-year medical student, masters in library science), previous formal MEDLINE training (no formal MEDLINE training, less than two hours, two to

four hours, more than four hours), and number of MEDLINE searches completed in the past six months. The Spearman *rho* correlation analysis shows a statistically significant relationship between total score and educational level (.516, $p<.001$, 2-tailed), total score and formal MEDLINE training (.441, $p<.001$, 2-tailed), and total score and number of searches performed in the past six months (.416, $p<.001$, 2-tailed).

Figure 8 compares the medians, quartiles, and extreme values of the total scores by educational level. Mean scores by educational level are reported in Table 9. Analysis of variance indicates that the mean scores by educational groups are not the same (see Table 10). Tukey's HSD multiple comparisons analysis shows significant differences (at the .05 level) between all educational groups).

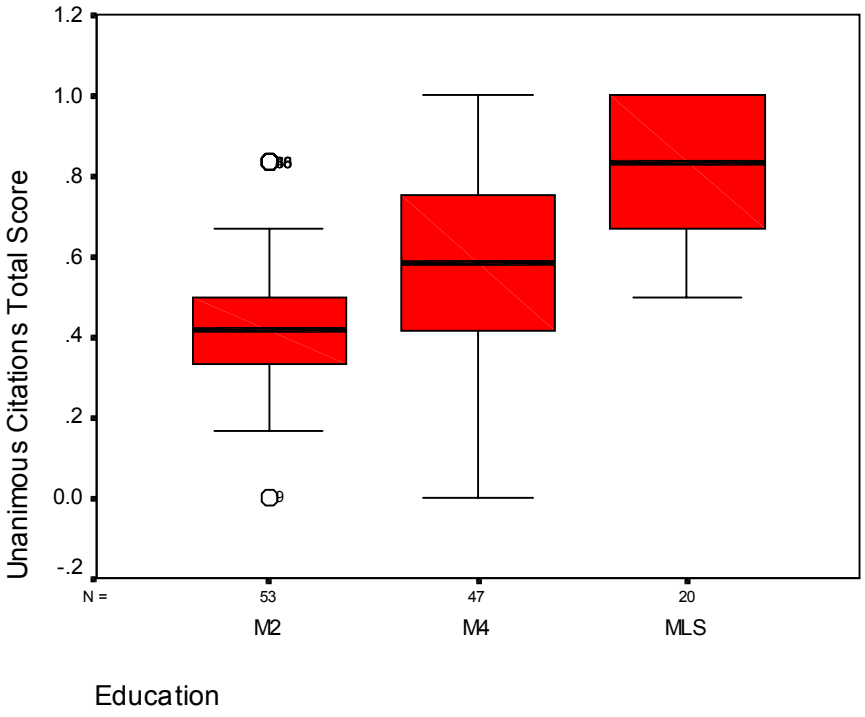

Figure 8. Medians, quartiles, and extreme values of total score by educational level.

Table 9

Mean Total Scores by Educational Level

Education	Mean	Number of Subjects	SD
M2	.4324	53	.1962
M4	.5674	47	.2189
MLS	.8000	20	.1822
Total	.5465	120	.2396

Table 10

ANOVA Summary: Mean Total Scores by Educational Level

	Sum of Squares	Df	Mean Square	F	Sig.	F_{cv}	Sig.
Between Groups	1.996	2	.998	24.150	.001	19.5	.05
Within Groups	4.835	117	4.132E-02				
Total	6.830	119					

ANOVA indicates that the mean scores are not the same for different levels of formal MEDLINE training (see Table 11). Tukey's HSD multiple comparisons analysis shows a statistically significant difference between the group with more than four hours of formal MEDLINE training and the other categories of training.

Table 11

ANOVA Summary: Mean Total Scores by Levels of Formal MEDLINE Training

	Sum of Squares	Df	Mean Square	F	Sig.	F_{cv}	Sig.
Between Groups	1.652	3	.551	12.348	.001	8.55	.05
Within Groups	5.086	114	4.461E-02				
Total	6.738	117					

There was a positive correlation between reported number of searches completed in the past six months and total score (Pearson correlation=.3421, $p<.001$, 2-tailed).

Scores for individual questions were also related to education level, but not always as one would have assumed. Table 10 shows the Spearman *rho* correlation coefficient for each of the questions and each educational level. All are statistically significant, at least at the .05 level. Note that there is a negative correlation of educational level and score for Question 4.

Table 12

Relationship Between Educational Level and Question Score

Question Number	Spearman *rho*	Significance, 2-tailed
Q1	.243	.007
Q2	.213	.019
Q3	.588	.000
Q4	-.184	.044
Q5	.325	.000
Q6	.347	.000

Total Score and Confidence in Retrieval

The average of the six (one for each question) ratings of agreement with the statement "I found highly relevant articles that I expect would answer the question asked" for each subject was considered a general indication of personal confidence in the ability to retrieve relevant articles. Overall, subjects were confident that they had found relevant articles for the test questions (mean=4.06, SD=.59, n=114). This indicator of confidence increased based on educational level (M2=mean of 3.88,

M3=mean of 4.03, and MLS=mean of 4.60). This difference was not statistically significant. For individuals, the average of ratings of confidence in retrieving relevant articles and the total score have a statistically significant and positive correlation (Pearson correlation=.398, $p<.001$, 2-tailed).

Appropriateness of Question Scores

Useful questions to ask about the test are "Do the test results appropriately categorize subjects by a known group membership?, and "Could individual question scores be used as predictors in cases where group membership is not known?" In this study, question scores were evaluated as predictors of mastery in MEDLINE searching.

Multinomial logistic regression is a robust method, requiring categorical dependent variables and continuous variable covariates. In this case, the categorical dependent variable is mastery represented by two educational levels, M2s (presumed unskilled searchers) and MLSs (presumed skilled searchers). The continuous variable covariates are individual question scores. The data fit the model of multinominal regression (chi-square=70.817, df=6, $p<.001$). The scores for M4s were not analyzed since they represent more heterogeneous skill levels.

Logistic regression produces likelihood ratio tests for the individual effects in the final model [87]. Significant chi-square statistics indicate the questions that are significantly related to group membership. According to this analysis, Questions 3, 6, and 2 are the better predictors of educational level group membership. Question 4 is flagged by SPSS with the diagnostic that "there may be quasi-complete separation in the data. . . . Some parameter estimates will tend to infinity," indicating that the inclusion of this question resulted in a mathematical processing problem. As noted

above, there was a negative correlation between educational level and Question 4. The likelihood ratios for the six questions are listed in Table 13.

Table 13

Likelihood Ratio Tests of Questions

Question	Chi-square	*p*
Q1	6.687	.010
Q2	15.889	.000
Q3	32.931	.000
Q4	1.185	.276
Q5	5.572	.018
Q6	20.379	.000

The classification table (Table 14) shows the observed vs. predicted group membership based on these questions. Question 4 was not calculated in the classification analysis due to mathematical processing problems noted above. The five-item test accurately classifies 94.5% of the subjects.

Table 14

Classification of Group Membership of Subjects Using 5 Question Scores

Group Membership		Predicted		Percent Correct
Observed		M2	MLS	
M2	53	51	2	96.2%
MLS	20	2	18	90.0%
Overall		72.6%	27.4%	94.5%

The probabilities generated by the logistic regression procedure can be used to determine the sensitivity and specificity of different cutoff scores for a test by means of receiver operating characteristic (ROC) analysis. Comparing the scores of presumed

unskilled searchers (M2s) to presumed skilled searchers (MLSs) results in the ROC curve given in Figure 9. The curve represents the accuracy of the five-item test (minus Question 4) in distinguishing unskilled from skilled searchers. The area under the curve is .93, which means that 93% of all possible M2-MLS pairs would be accurately classified as unskilled searchers or skilled searchers. This 93% accuracy was statistically significant (SE=.027, *p*=.001). The cutoff point closest to the upper left-hand corner indicates the cutoff point (score) at which both sensitivity and specificity are optimized. In this case, using scores greater than .50, 70% of the skilled searchers are correctly classified and 7.5% of unskilled searchers are incorrectly identified as skilled (false positives). At this cutoff score, the test is said to have .70 sensitivity and .925 specificity.

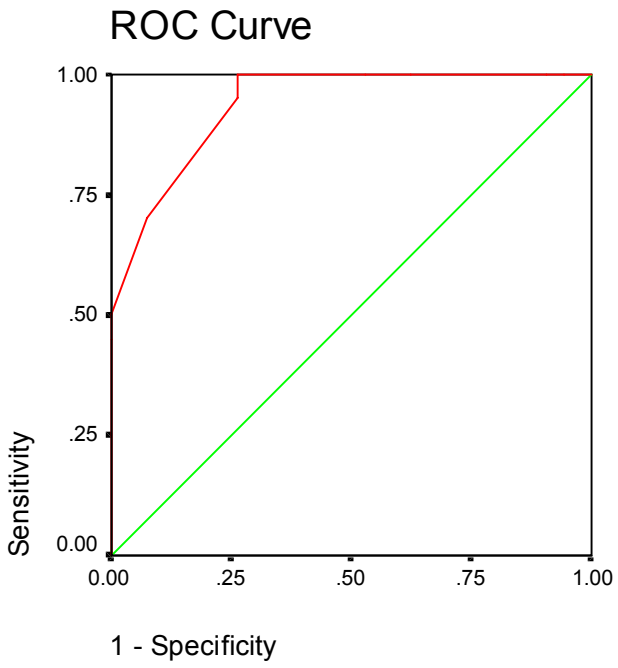

Figure 9. ROC curve for 5-item test, unskilled searchers (M2s) and skilled searchers (MLSs).

Test Reliability

The reliability coefficient alpha for the six individual test scores was .3843. Question 4 correlates negatively with Questions 1, 3, and 6. As noted above, there is also a negative correlation with educational level. When Question 4 is eliminated, the coefficient alpha is .4775.

The standard error of measurement for a test with .4775 reliability coefficient and a standard deviation of the test score of .27 is .19. This means that there is approximately a one in three chance that a person's observed test score differs from that person's "true" test score by as many as .19 points in either direction [88]. Also, there is a 1 in 20 chance that the true score differs by as much as .38 points in either direction (two standard errors from the observed score).

The Spearman-Brown Prophecy Formula provides a way of calculating how much longer a test should be to achieve a given level of reliability. According to this formula (see Chapter 3), the test would have to be five times longer (25 questions) to achieve a reliability coefficient of .82.

Time Spent Taking the Test

The test was administered under "field" conditions. Especially in the classroom administration (M2s), tracking time was difficult, although no one took more than two hours for the entire test. One test proctor kept careful track of the time it took subjects to complete their searches. Table 15 reports the average, range, and standard deviation for the amount of time spent on each search as well as for the entire test. It should be noted that all of the subjects for whom time was recorded were M4s. The average time to

complete one question on the test for this subset of subjects was approximately 13 minutes, or 78 minutes for the entire test.

Table 15

Average Time to Complete Each Question and Entire Test for M4s (n=19)

	Time			
Question Number	Minimum	Maximum	Mean	SD
Q1	3.0	20.0	13.316	5.012
Q2	4.5	20.0	10.711	4.395
Q3	7.0	20.0	15.526	4.647
Q4	4.0	20.0	10.895	4.471
Q5	3.0	20.0	11.316	4.726
Q6	8.0	20.0	16.000	4.177
Total Test	7.7	17.0	12.961	2.649

Comments

Subjects' comments about taking the test varied considerably. Selected comments are listed in Figure 10.

Comments
This has helped me.
It was straight-forward.
Was a good exercise and a learning experience.
This is very difficult.
I am getting better as I go on, I believe.
I believe that experience using MEDLINE is the best way to learn it.
I am learning new tricks on each search.
I am surprised such data exists.
Learned something (author/title search).
This is a complete waste of time.
Productive, but too much time was allotted for activity.
Even though this session was useful, I think 2 searches would be sufficient.
This should not take 2 hours.
Getting tired.
Growing old an hour ago.

Figure 10. Representative comments about the test in general.

CHAPTER 5

CONCLUSIONS AND SUMMARY

The goal of this research was to develop and evaluate a method for assessing medical students' information retrieval skills. Following a documented procedure for test development, experts formulated six search scenarios and questions. One hundred and forty-five subjects, representing a wide range of MEDLINE search expertise, completed the six-item test under timed conditions. Data related to educational level, search training, search experience, confidence in retrieval, difficulty of search, and score were collected and analyzed. The underlying question addressed in this chapter is "Does the method proposed identify the information retrieval skills of medical students?" More specifically, is the MEDLINE Metric method—not just the test—valid, reliable, and feasible to use in the medical curriculum? The discussion is organized around the eight research questions proposed. This chapter concludes with comments about feasibility and recommendations for further research.

Research Questions

1. Is the proposed method of assessing search effectiveness valid?

Evidence to support the construct validity of the overall method (as illustrated in Figure 2) comes from comparing it with other methods that have attempted to accomplish the same task. This method builds on previously reported research that used questions based on clinical scenarios in a test environment and compared students' retrieval with that of librarian search experts.

Evidence of content validity for the test includes the agreement by experts about the skills and knowledge necessary to successfully retrieve information relevant to a clinical question from the MEDLINE database. The scenarios and questions in this study were specifically designed for search skill assessment. Other studies used either spontaneous clinical questions or questions designed for a clinical course exam.

The assessment is based on retrieval, not strategy—search results rather than process. The skill directly measured is the ability to retrieve the citations search specialists and content experts identified as highly relevant. Thorndike states that "the correspondence between the test blueprint (i.e., explicit statement of what a test is intended to measure) and the definition of the trait to be measured *is* the content validity of the test"[89].

2. Does the test discriminate among different performance levels?

A criterion-referenced test should distinguish between those who have achieved mastery and those who have not. In this study, the purposive selection of subjects included people with a variety of assumed MEDLINE search experience and skills, from unskilled to skilled searchers. Test scores were analyzed to determine if the assessment distinguished among presumed levels of skill. Mean and median test scores differed by level of education and self-reported amount of previous training. Logistic regression analysis showed that the test appropriately classifies unskilled and skilled searchers.

The fact that the test does a good job of discriminating between different performance levels also supports evidence of the test's validity.

3. Are the results of the test reliable?

Information about the reliability of the test itself comes from different sources. Test administration in the study occurred under "field" conditions, similar to how the test would probably be administered in practice. Grading the test was an objective process, not dependent on any judgements that might vary from person to person. Subjects' results were compared to the list of key citations. The test appears to be a valid measure of search skills in that it discriminates unskilled from skilled searchers, indicating some level of reliability. Cronbach's coefficient alpha is low (.4775), but this may not be a good measure of reliability for a criterion-referenced test [90]. A carefully designed test-retest study of fourth-year medical students is the logical next step to assess reliability.

4. How many test searches are needed to reasonably assess a student's performance level?

With Cronbach's coefficient alpha and the number of items in a test, one can use the Spearman-Brown Prophecy formula to calculate the effect of increasing the number of items on the coefficient alpha for the test. Based on this calculation, the test would have to be five times as long—include twenty-five search questions—to have a reliability coefficient of .82. If the average search takes 13 minutes, the test would be approximately 5.5 hours long.

5. Is the method of assessment reliable across different MEDLINE search systems?

There was not enough data for this comparison. The majority (97) of subjects used the Ovid Web MEDLINE interface. The next most-used system was NLM's PubMed. Of the nine people who used PubMed, seven were MLSs, a disproportionate number of skilled searchers.

6. What is the relationship between level of education and test score?

There is a statistically significant, positive relationship between level of education and test score. This is true for both the original, six-search test (Spearman *rho*=.516, $p<.001$, 2-tailed) and for the test with five searches (Spearman *rho*=.604, $p<.001$, 2-tailed).

7. What is the relationship between previous MEDLINE search training and test score?

There is a statistically significant, positive relationship between (self-reported) previous MEDLINE search training and test score. This is true for both the original, six-search test (Spearman *rho*=.441, $p<.001$, 2-tailed) and for the test with five searches (Spearman *rho*=.490, $p<.001$, 2-tailed).

8. What is the relationship between previous MEDLINE search experience and test score?

There is a statistically significant, positive relationship between previous MEDLINE search experience (determined by self-reported number of MEDLINE searches completed in the past six months) and test score. This is true for both the original, six-search test (Spearman *rho*=.416, $p<.001$, 2-tailed) and for the test with five searches (Spearman *rho*=.490, $p<.001$, 2-tailed).

Feasibility of test construction and test administration was also considered. The construction of the test includes identifying current medical topics and writing scenarios and questions appropriate for medical students, as well as coming to consensus about key citations. The original expectation was that the medical content experts would be more involved in constructing the scenarios, but it was difficult enough to get them to identify

questions, due primarily to their lack of time for this activity. Also, it became apparent that each scenario and question required at least two iterations of editing to result in consensus around retrieval. The search specialists could easily suggest modifications in wording that might affect retrieval. The selection by content experts of the same key citations that the searchers agreed on indicates that searchers can construct reasonable medical search scenarios and retrieve highly relevant key citations. The role of the content expert could be limited to suggesting topics and verifying key citations.

Using realistic questions about current medical topics, the "live" MEDLINE database, and allowing subjects their choice in search systems lends authenticity and ecological validity to the test. It also introduces the possibility that "better" key citations will become available over time and may appear first in one MEDLINE system and not another. This in fact happened with the question about Calan and heart disease. Two expert searchers used a version of MEDLINE that was more current by just a few weeks, and they found a highly relevant article that the other searchers, using a less frequently updated version of MEDLINE, did not retrieve. At the other extreme, as MEDLINE files are updated, they get segmented into groups of years. The so-called "current file" was 1995 to present when this study began in 1998. With the change in calendar year to 1999, the current file changed to 1996 to present. One of two key citations for Question 3 was no longer retrievable from the current file. These problems can be addressed by giving more explicit instructions about what years to search and, in the case of new, more relevant citations, either by continuously checking the database or by giving a broader range for retrieval.

Checking the database continuously would be burdensome. Allowing people to choose more citations, perhaps as many as ten, is an easier way to keep key citations the same over a reasonable period of time. Experience with this test suggests other good reasons for increasing the number of citations selected. In phase one of the study, subjects were instructed to select a certain number of highly relevant articles, and that number was not the same for each of the questions. Some people selected more citations than they were supposed to for some questions, perhaps because the instructions were not the same. Scores for those questions were coded as "included key citation but selected more" and processed as missing values. In phase two of the study, subjects were instructed to select between two and five citations for all questions. Recoding the first-phase data using this more liberal criterion converted 31 scores in 25 cases to valid scores. Further support for letting people select more articles was the feedback of subjects who were experienced searchers. More than one felt compelled to select more articles than asked for despite the instructions. Some made comments like: "I would be much happier if I could give the requester the following 4 citations" and "Gave 3 citations even though 2 were asked for because I thought the most recent one by the current NEJM editors updating Relman's writings would be essential." Future administration of the test should allow people to select more citations.

The 20-minute per search time limit was not a problem, but a test that lasted more than two hours to assess just MEDLINE skills might not be acceptable to medical schools. As suggested by others, the Objective Structured Clinical Exam context would be appropriate for search skills assessment since OSCEs focus on performance, use clinical scenarios, and are timed tests.

Recommendations for Future Studies

The next study should focus on assessing the reliability of this test. A sample of fourth-year medical students should take the test and then retake it two weeks later, without intervening instruction.

Another study should assess the reliability of the method for constructing scenarios and test questions, and for identifying key citations. Other search specialists and medical content experts would be asked to follow the process described in Figure 2.

Using the same questions, other subjects should take the test with the instructions to select more citations to determine the effect of a larger retrieval set on test scores. The hypothesis would be that more subjects would do better on the test, but the test would still discriminate well between unskilled and skilled searchers.

An interesting study could test the hypothesis that a search system with more advanced features, such as PubMed, would result in higher scores even among novices.

The MEDLINE Metric method might be tested in other health science domains, such as nursing or public health. Content experts, scenarios, and questions would have to be related to those subject areas.

Summary

Increasingly, medical students are expected to master informatics skills such as information retrieval. Performance-based assessment is more appropriate than tests of factual recall to measure these skills. This study offers a method for constructing and administering a performance-based test to identify mastery in searching the MEDLINE database.

APPENDIX

APPENDIX A

MEMO TO CONTENT EXPERTS REQUESTING

CASE SCENARIOS

I am writing to follow up with you about developing case scenarios for a test that will assess medical students' MEDLINE search skills. This is a project that Rachel Bramson and I are working on, and it will also be used for my dissertation research. Basically, we need scenarios that would result in a search of the journal literature. These are some examples of scenarios I have used for the Internal Medicine OSCE at Texas A&M; they will not be used for the final test:

1. One of your patients with adult onset diabetes brings you an article about the Diabetes Control and Complications Trial (DCCT) that has shown that intensive insulin therapy ("tight control") will prevent or delay the onset of retinopathy, nephropathy, and neuropathy in Type I diabetes. He wants to know if there is evidence that such therapy would be of benefit to him.

2. Alzheimer's disease accounts for approximately two-thirds of all cases of dementia in the US and $90 billion in health care costs annually. You have been asked by your primary care group practice colleagues to organize the next journal club and to select articles that will update them on the current status of the treatment for Alzheimer's disease.

3. A 46-year-old woman with arthritis says that she read an article in *Prevention* magazine about the positive effects of fasting. She is curious about this and about other diet-related therapies. She asks if there has been "legitimate" research in this area. You offer to send her some articles.

We want to use "core" topics, familiar to third-year medical students, for the scenarios. The Society of General Internal Medicine and the Clerkship Directors in Internal Medicine have developed a Core Medicine Clerkship Curriculum recommended for third-year medical students. Suggested training problems include:

Health promotion, disease prevention, screening
Cough
Dysuria
Back pain
Altered mental status
Joint pain
Chest pain
Abdominal pain
Abnormal fluid or electrolyte finding
Chronic obstructive pulmonary disease
HIV infection
Congestive Heart Failure
Diabetes mellitus
Dyslipidemias
Substance abuse
Smoking
Common cancers
Anemia
Hypertension

Would you please develop three scenarios based on conditions from this list? They need not be long. Please indicate the audience, e.g., patients for example scenarios #1 and #3, primary care physicians for scenario #2. If you have an idea of a topic that would be good

to search because it is interesting or controversial but you can't think of a question, please send it on. These scenarios and ideas are to be the basis of very explicit search questions developed by search experts to test specific search skills. I am relying on you, as content experts in medicine, to identify "hot" topics.

We'd like to have these scenarios in two weeks, if possible. You may e-mail them to g-hannigan@tamu.edu. Please contact me if you have questions—and, thanks.

APPENDIX B

LETTER TO SEARCH SPECIALISTS

June 23, 1998

TO:
FROM: Gale G. Hannigan
SUBJECT: Search Specialist Role

Thank you for agreeing to participate as a search specialist in my dissertation research to develop a method to assess medical students' MEDLINE search skills. I have had the chance to talk with each one of you and look forward to working with you.

The search specialists in this project have the very important tasks of defining educational objectives that should be assessed and developing the search questions for the test. As one of the members of my dissertation committee said, the test is worthless unless the questions themselves are valid and measure what they are supposed to measure.

With regard to objectives, I want to point out that this method of assessment is designed to look at results rather than process. Many researchers have studied search logs but, in this case, we will measure students' ability to retrieve the most relevant articles and not consider how they go about doing their searches. Even so, the assumption is that the students will not get good results unless they have specific search skills. Please help me identify these skills by adding to the following list:

MEDLINE Information Retrieval Skills Educational Objectives

Students will demonstrate the ability to:

1. Use appropriate MeSH terms
2. Limit search retrieval by language, year, age groups, human subjects, publication type
3. Use appropriate subheadings to limit retrieval
4. Combine concepts using Boolean OR and AND operators appropriately
5. Retrieve a specific citation given information about the title, author, year
6. Select articles appropriate for a defined audience (e.g., peers, patients)

I have based this list on personal experience, training materials and publications (e.g., NLM training materials, syllabi, books). Please review and comment, and add to it

considering that these objectives should be the basis of training to prepare students to take our test (which I am calling MEDLINE Metric).

The educational objectives that we agree on will obviously also influence the questions we develop. Each of you is working to develop three to five search questions based on topics provided by content experts. As you do this, please consider and identify which of the objectives each question addresses. When we have developed our questions and objectives, we should be able to complete the following matrix:

	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
Objective						
Objective						
Objective						
Objective						
Objective						
Objective						
Objective						
Objective						

How well the questions cover the objectives and complement one another will help us select the questions to use on the test.

I have enclosed for your reading pleasure (ha!) a copy of my proposal, with the revisions suggested by my doctoral committee members. You will probably find Chapter 3 (Methods) useful for understanding how I think this is all going to work. Please, if you have comments or criticisms, let me know them. Better now than later! My e-mail address is g-hannigan@tamu.edu

I hope that we can define objectives and develop questions by the end of July and then test in August and September. It's a pretty ambitious schedule since we four have to develop and try out the test questions. If you can get me your thoughts about objectives by July 8 and your questions and the objectives they cover by mid-July, that would be wonderful. Please let me know if this is not possible.

Thanks again for contributing to this project. If we can develop a valid method to assess students' MEDLINE skills, it will be an important contribution to medical informatics education. Of course, your efforts will be acknowledged in the dissertation and any subsequent publications. Your participation is probably "CV'able" – the project is funded, in part, by the ISI/MLA Doctoral Fellowship I received in 1996.

cc: A Cleveland

APPENDIX C

SAMPLE TEST PACKET

MEDLINE Metric
Informed Consent Form for Paid Volunteers

I will be asked to provide information about previous MEDLINE search training and experience and to complete six searches on medical topics.

This assessment benefits me by providing search practice and feedback about my current search skills. I understand that I will be paid $15.00 upon completion of the searches and related paperwork as compensation for time, which is approximately 2.5 hours. Approximately 120 people will participate in this study. I understand that I may stop my participation at any time, but that I will not be compensated if I do not complete the searches and paperwork. Scores will be reported back to me for my information; they will not be used for any course grade. Data from this activity will be used as part of an educational research project to develop a valid method for assessing MEDLINE search skills. All information that identifies me will be separated from the data before final analysis.

THIS PROJECT HAS BEEN REVIEWED BY THE UNIVERSITY OF NORTH TEXAS COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS (940-565-3940) AND THE INSTITUTIONAL REVIEW BOARD—HUMAN SUBJECTS IN RESEARCH, TEXAS A&M UNIVERSITY. For research-related problems or questions regarding subjects' rights, the Institutional Review Board may be contacted through Dr. Richard E. Miller, IRB Coordinator, Office of Vice President for Research and Associate Provost for Graduate Studies at (409) 845-1811.

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study.

I have been given a copy of this consent form.

_____	_____	_____
Date	Participant's signature	Last name printed

_____
Signature of Principal Investigator

If you have questions about this study, contact: Gale G. Hannigan, Texas A&M College of Medicine Learning Resources, College Station, TX; (409) 845-0514 or Dr. Rachel Bramson, Department of Community and Family Medicine, Texas A&M College of Medicine, College Station, TX; (409) 845-7829.

MEDLINE Metric
Informed Consent Form for Paid Volunteers

I will be asked to provide information about previous MEDLINE search training and experience and to complete six searches on medical topics.

This assessment benefits me by providing search practice and feedback about my current search skills. I understand that I will be paid $15.00 upon completion of the searches and related paperwork as compensation for time, which is approximately 2.5 hours. Approximately 120 people will participate in this study. I understand that I may stop my participation at any time, but that I will not be compensated if I do not complete the searches and paperwork. Scores will be reported back to me for my information; they will not be used for any course grade. Data from this activity will be used as part of an educational research project to develop a valid method for assessing MEDLINE search skills. All information that identifies me will be separated from the data before final analysis.

THIS PROJECT HAS BEEN REVIEWED BY THE UNIVERSITY OF NORTH TEXAS COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS (940-565-3940) AND THE INSTITUTIONAL REVIEW BOARD—HUMAN SUBJECTS IN RESEARCH, TEXAS A&M UNIVERSITY. For research-related problems or questions regarding subjects' rights, the Institutional Review Board may be contacted through Dr. Richard E. Miller, IRB Coordinator, Office of Vice President for Research and Associate Provost for Graduate Studies at (409) 845-1811.

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study.

I have been given a copy of this consent form.

_____	_____	_____
Date	Participant's signature	Last name printed

_____
Signature of Principal Investigator

If you have questions about this study, contact: Gale G. Hannigan, Texas A&M College of Medicine Learning Resources, College Station, TX; (409) 845-0514 or Dr. Rachel Bramson, Department of Community and Family Medicine, Texas A&M College of Medicine, College Station, TX; (409) 845-7829.

SUBJECT'S COPY

Pre-Search Questionnaire

Name (please print) _____

1. Educational Level (check one):

___ Undergraduate

___ First-year medical student

___ Second-year medical student

___ Third-year medical student

___ Fourth-year medical student

___ Resident Physician

___ Masters in Library Science

2. Number of hours of formal MEDLINE search skills training – class, session, workshop taught by a MEDLINE expert (check one):

___ No formal MEDLINE training

___ Less than 2 hours of formal MEDLINE training

___ 2-4 hours of formal MEDLINE training

___ More than 4 hours of formal MEDLINE training

3. Approximately how many MEDLINE searches have you done in the past six months?

_____ searches

4. Which search system will you be using today? (Ask if you are not sure)

___Ovid – Web interface

___Ovid – Telnet interface

___Ovid – CD (at Scott & White Library)

___PubMed

___Other, please specify _____

1. You are a patient with chronic heart disease who regularly takes Calan. A co-worker tells you he has heard that some of the drugs used for heart disease increase the risk of getting cancer. Should you be concerned? Find studies that address this. Use the most current file of MEDLINE (1995-1998). Select 3 articles that seem most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

2. You are an ER physician and next week you meet with hospital administrators to propose the use of CT scans for patients suspected of having appendicitis. You think that this diagnostic test would reduce the number of unnecessary appendectomies and reduce hospital costs. You want to find articles to cite to support your proposal. Use the most current file of MEDLINE (1995-1998). Select 3 articles that seem most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

3. You are a busy first-year pediatrics resident and just admitted a child with a GI bleed. You wonder if it could have been caused by her use of NSAIDs (non-steroidal anti-inflammatory agents). Find current articles that deal specifically with children under twelve; you don't have time to read case reports. Use the most current file of MEDLINE (1995-1998). Select 3 articles that seem most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

4. Diabetes mellitus is a significant public health problem in the US. You have been invited to give a lecture to medical students about diabetes and you want to start with current statistics about the direct and indirect medical costs associated with this disease. Use the most current file of MEDLINE (1995-1998). Select 3 articles that seem most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

5. You are a pediatrician and see lots of middle ear infections. For patients with recurrent infections, you wonder about the effectiveness of using antibiotics, specifically amoxicillin, for prevention. You worry about the over use of antibiotics; is it better not to use them to prevent these infections? Find articles that provide relevant data. Use the most current file of MEDLINE (1995-1998). Select 2 articles that are most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

6. Your attending mentions the "Ingelfinger policy," which you have never heard of. She says that, back in the 1980s, Arnold Relman, editor of the *New England Journal of Medicine* wrote about it, and asks you to find the articles. Select 2 articles that seem most relevant.

You have 20 minutes to:

a. Complete your search and the attached Post-Search Questionnaire.
b. Print out the specified number of citations and your search strategy. If you print more, make sure to circle the specified number of articles that best answer the question.
c. To print, after you select the citations you want, go to the bottom of the page to the **Citation Manager** screen. Check the box that says **Include Search History** (in the Citation Format column). This default will prevail for future searches, unless the system goes down. Click on **Display**, and the selected citations will display along with your search strategy. Go to **File**, then **Print**, to print out your search and strategy.

When you return to the **Main Search Page**, if you want to clear out your statements to start a new search (you don't have to do this), type: **pg 1 — (the number of your last statement)**. Hit the enter key.

Post-Search Questionnaire

Please circle the number that best indicates your agreement with the following statements about this search:

1. I found highly relevant articles that I expect would answer the question asked.

Strongly disagree				Strongly agree
1	2	3	4	5

2. Did you have any difficulty with this search?

___ No

___ Yes If Yes, please explain:

3. Comments about the search or this activity:

APPENDIX D

FEEDBACK FORM

Name _____
Score on MEDLINE assessment _____

Your score is based on your retrieval compared to the consensus retrieval of four expert searchers and the selection of the most relevant articles by a clinician. While you may have found other relevant articles, if your score is low, you did not identify the "highly retrievable, highly relevant" articles these experts identified. Remember, the score is for your self-assessment, and not part of your course grade.

In general, the strategy that would <u>most</u> improve everyone's searches would be to search one concept at a time; use ANDs and ORs to combine search statements, not search terms. For example,

1. Heart Disease		1. Heart Disease AND Diet
2. Diet	NOT	
3. 1 AND 2		

By using this building blocks approach (one concept at a time), you can take advantage of automatic system features such as mapping to the MEDLINE subject headings and subheadings.

Other suggestions for improving your retrieval, based on your strategies are:

___ Identify <u>all</u> concepts in the question (e.g., year, age, diseases, drugs)

___ When searching and selecting, keep in mind the intended audience for the information (patients, subspecialists, administrators)

___ Consider the quality of the source; letters, case studies, articles in state medical journals have less weight than research studies in major journals

___ [Brackets around a title indicate the article is not in English]

___ When selecting articles, look for terms specific to your question, "acute" vs. "chronic," health care in a specified country, etc.

___ Unlike the World Wide Web, MEDLINE searches do not cumulate. In the search:
1. Heart Disease
2. Diet

#2 includes <u>all</u> articles on diet, not only those about heart disease and diet.

___ MEDLINE is a disease-oriented database. When possible, start with the disease term.

Questions? Contact Gale Hannigan at 771-5443 or g-hannigan@tamu.edu

m3m4 version

APPENDIX E

CODING FORM

ID _____

EDCU
0 – Undergrad
1 – First-year medical student
2 – Second-year medical student
3 – Third-year medical student
4 – Fourth-year medical student
5 – Resident physician
6 – Masters in Library Science

TRAIN
0 – No formal MEDLINE training
1 – Less than 2 hours formal MEDLINE training
2 – 2-4 hours formal MEDLINE training
3 – More than 4 hours formal MEDLINE training

SEARCHES ____ searches done in past six months

SYSTEM
1 – Ovid – Web interface
2 – Ovid- Telnet
3 – Ovid – CD (S&W library)
4 – PubMed
5 – Other

TESTVAR ____(variation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 written lower corner of question 1)

9-missing	1-No 2-Yes 9-Missing	1-Comment 2-None	See below
REL1 ____	DIFF1 ____	COMMENT1____	SCORE1 ____ (0,1,8,9)
REL2 ____	DIFF2 ____	COMMENT2____	SCORE2 _____(0,1,8,9)
REL3 ____	DIFF3 ____	COMMENT3____	SCORE3 _____(0,.5,1,8,9)
REL4 ____	DIFF4 ____	COMMENT4____	SCORE4 _____(0,1,8,9)
REL5 ____	DIFF5 ____	COMMENT5____	SCORE5 _____(0,1,8,9)
REL6 ____	DIFF6 ____	COMMENT6____	SCORE6 _____(0,1,.5,8,9)

0 – did not retrieve any of the answer citations
1 – full credit: retrieved 1 and there was only 1, retrieved 2 and there were 2 in answer
.5- retrieved 1 and there were 2 citations in answer
8- retrieved answer citation(s) but had more than question asked for
9- no information, missing data

REFERENCES

1. Swanson AG, Anderson MB. Educating medical students. Assessing change in medical education—the road to implementation. Acad Med 1993 Jun;68(6 Suppl):S37.
2. Lundberg GD. Perspective from the editor of JAMA, The Journal of the American Medical Association. Bull Med Libr Assoc 1992 Apr;80(2):110.
3. Council on Medical Education. American Medical Association. Future directions for medical education: a report of the Council on Medical Education, adopted June 15, 1982 by the House of Delegates of the American Medical Association. Chicago, IL: The Association, 1982.
4. The new biology and medical education: merging the biological, information, and cognitive sciences. Report of a conference sponsored jointly by the University of North Carolina and the Josiah Macy, Jr. Foundation. New York, NY: The Foundation, 1983.
5. Panel on the General Professional Education of the Physician and College Preparation for Medicine. Association of American Medical Colleges. Physicians for the twenty-first century: the GPEP report. Washington, DC: Association of American Medical Colleges, 1984.
6. Gastel B, Rogers DE, eds. Clinical education and the doctor of tomorrow. Proceedings of the Josiah Macy, Jr. Foundation National Seminar on Medical Education. Adapting clinical medical education to the needs of today and

tomorrow, June 15-18, 1988. New York, NY: New York Academy of Medicine, 1989.

7. Panel on the General Professional Education of the Physician and College Preparation for Medicine. Association of American Medical Colleges, op. cit.
8. Swanson AG, Anderson MB, op. cit.
9. Swanson AG, Anderson MB, op. cit., S9-10.
10. Swanson AG, Anderson MB, op. cit., S38.
11. The Medical School Objectives Project Writing Group. Learning objectives for medical student education. Guidelines for medical schools: report I of the Medical School Objectives Project. Acad Med 1999 Jan;74(1):13.
12. Ibid., 17.
13. Shortliffe EH, Perrault LE. Medical informatics: computer applications in health care. Reading, MA: Addison-Wesley, 1990:20.
14. Association of American Medical Colleges. Medical School Objectives Project: Medical informatics objectives. [Web document]. Washington, DC: The Association, 1998. [cited 7 Feb 1999] Available from Internet: <http://www.aamc.org/meded/msop/informat.htm>.
15. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 5th ed. New York, NY: Macmillan, 1991;53.
16. Saracevic T, Kantor P. A study of information seeking and retrieving: III. Searchers, searches, and overlap. JASIS 1988 May;39(3):204.

17. McKibbon KA, Haynes RB, Dilks CJ, Ramsden MF, Ryan NC, Baker L, Flemming T, Fitzgerald D. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comput Biomed Res 1990 Dec;23:592.
18. Stross JK, Harlan WR. The dissemination of new medical information. JAMA 1979 Jun 15;241(24):2662-4.
19. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts, Treatments for myocardial infarction. JAMA 1992 Jul 8;268(2):240-8.
20. Williamson JW, German PS, Weiss R, Skinner EA, Bowes F. Health science information management and continuing education of physicians: a survey of U.S. primary care practitioners and their opinion leaders. Ann Intern Med 1989 Jan 15;110(2):151-60.
21. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985 Oct;103(4):596-9.
22. Connelly DP, Rich EC, Curley SP, Kelly JT. Knowledge resource preferences of family physicians. J Fam Pract 1990 Mar;30(3):353-9.
23. Florance V. Medical knowledge for clinical problem solving: a structural analysis of clinical questions. Bull Med Libr Assoc 1992 Apr;80(2):140-9.
24. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. Ann Intern Med 1990 Jul 1;113(1):69-76.
25. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med 1987 Aug;107(2):224-33.

26. Institute of Medicine (U.S.) Committee on Clinical Practice Guidelines. Guidelines for clinical practice: from development to use. Washington, DC: National Academy Press, 1992.

27. Guyatt GH, Rennie D. Users' guides to the medical literature (editorial). JAMA 1993 Nov 3;270(17):2096-7.

28. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (editorial). BMJ 1996 Jan 13; 312(7023):71.

29. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. JAMA 1993;270(17):2094.

30. U.S. National Library of Medicine. MEDLINE. [Web document]. Bethesda, MD: National Institutes of Health, 1999. [rev. 14 Sept 1999] Available from Internet: http://www.nlm.nih.gov/pubs/factsheets/medline.html.

31. Billings JS. Prospectus. Index Medicus, Vol 1. New York, NY: F. Leypoldt, 1879.

32. Ibid.

33. Welborn V, Kuehn JJ. End-user programs in medical school libraries: a survey. Bull Med Libr Assoc 1988 Apr;76(2):137-40.

34. Salisbury L, Toombs HS, Kelly EA, Crawford S. The effect of end-user searching on reference services: experience with MEDLINE and Current Contents. Bull Med Libr Assoc 1990 Apr;78(2):188-91.

35. Haynes, RB, Walker CJ, McKibbon KA, Johnston ME, Willan, AR. Performances of 27 MEDLINE systems tested by searches with clinical questions. J Am Med Inform Assoc 1994 May-Jun;1(3):285-295.

36. Association of Academic Health Sciences Library Directors. Annual statistics of medical school libraries in the United Stated and Canada. 20th ed. Seattle, WA: The Association, 1998.

37. Wilson SR, Starr-Schneidkraut N, Cooper MD. Use of the Critical Incident Technique to evaluate the impact of MEDLINE. Final report, September 30, 1989. (Contract No. N01-LM-8-3529). [Wed document]. [cited 17 Apr 1998] Available from Internet: <http://www.nlm.nih.gov/od/ope/citackn.txt>.

38. Lindberg DA, Siegel ER, Rapp BA, Wallingford KT, Wilson SR. Use of MEDLINE by physicians for clinical problem solving. JAMA 1993 Jun 23-30;269(24):3124-9.

39. Chambliss ML, Conley J. Answering clinical questions. J Fam Pract 1996 Aug;43(2):140-4.

40. Scura G, Davidoff F. Case-related use of the medical literature. Clinical librarian services for improving patient care. JAMA 1981 Jan;245(1):50-2.

41. Klein MS, Ross FV, Adams DL, Gilbert CM. Effect of online literature searching on length of stay and patient care costs. Acad Med 1994 Jun;69(6):489-95.

42. Boyce BR, Meadow CT, Kraft DH. Measurement in information science. San Diego, CA: Academic Press, 1994;177.

43. Ibid., 180.

44. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. 2d ed. New York, NY: Springer-Verlag, 1995;118-9.

45. Fenichel CH. The process of searching online bibliographic databases: a review of research. Library Research 1980-81;2:107-27.

46. Fenichel CH. "Online information retrieval: identification of measures that discriminate among users with different levels and types of experience," Unpublished dissertation, Drexel University, 1979, 278p.
47. Saracevic T, Kantor P. A study of information seeking and retrieving: II. Users, questions, and effectiveness. JASIS 1988;39(3):177-196.
48. Poisson EH. End-user searching in medicine. Bull Med Libr Assoc 1986 Oct;74(4):293-9.
49. Wildemuth BM, Moore ME. End-user search behaviors and their relationship to search effectiveness. Bull Med Libr Assoc 1995 Jul;83(3):294-304.
50. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings, A study of use and usefulness. Ann Intern Med 1990 Jan 1;112:78-84.
51. McKibbon, op. cit.
52. Walker CJ, McKibbon KA, Haynes RB, Johnston ME. Performance appraisal of online MEDLINE access routes, Proc Annu Symp Comput Appl Med Care 1992;483-7.
53. Haynes RB, Ramsden MF, McKibbon KA, Walker CJ. Online access to MEDLINE in clinical settings: impact of user fees. Bull Med Libr Assoc 1991 Oct;79(4):377-81.
54. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL. Factors affecting students' use of MEDLINE. Comput Biomed Res 1993 Dec;26(6):541-55.

55. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, Shipman BL, McQuillan M. Effect of search experience on sustained MEDLINE usage by students. Acad Med 1994 Nov;69(11):914-920.

56. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL, op.cit., 546.

57. Shelstad KR, Clevenger FW. On-line search strategies of third year medical students; perception vs fact. J Surg Res 1994 Apr;56(4):338-44.

58. Ibid., 340.

59. Burrows SC, Tylman V. Evaluating medical student searches of MEDLINE for evidence-based information: process and application of results. Bull Med Libr Assoc 1999 Oct;87(4):471-6.

60. Cantor JC, Cohen AB, Barker DC, Shuster AL, Reynolds RC. Medical educators' views on medical education reform. JAMA 1991 Feb 27;265(8):1002-6.

61. Barrows HS, Tamblyn RM. Problem-based learning: an approach to medical education. New York, NY: Springer, 1980.

62. Albanese MA, Mitchell S. Problem-based learning: a review of literature on its outcomes and implementation issues. Acad Med 1993 Jan;68(1):52-81.

63. Berkson L. Problem-based learning: have expectations been met? Acad Med 1993 Oct;68(10 Suppl):S79-S88.

64. Small PA, Stevens CB, Duerson MC. Issues in medical education: basic problems and potential solutions. Acad Med 1993 Oct;68(10 Suppl):S89-S98.

65. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979 Jan;13(1):41-54.

66. Reznick R, Smee S, Rothman A, Chalmers A, Swanson D, Dufresne L, Lacombe G, Baumber J, Poldre P, Levasseur L, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. Acad Med 1992 Aug;67(8):487-94.

67. Woods SE, Francis BW. MEDLINE as a component of the objective structured clinical examination: the next step in curriculum integration. Bull Med Libr Assoc 1996 Jan;84(1):108-9.

68. Hannigan GG. MEDLINE OCSCE station for an internal medicine clerkship. Paper presented at the annual meeting of the MLA/South Central Chapter, Albuquerque, NM, Oct 5, 1997.

69. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. Acad Med 1996 Jan;71(1 Suppl):S19-S21.

70. Travis TA, Colliver JA, Robbs RS, Barnhart AJ, Barrows HS, Giannone L, Henkle JQ, Kelly DP, Nichols-Johnson V, Rabinovitch S, Ramsey DE, Riseman J, Rockey PH, Ross DS, Schrage JP, Steward DE. Validity of a simple approach to scoring and standard setting for standardized-patient cases in an examination of clinical competence. Acad Med 1996 Jan;71(1 Suppl):S84-6.

71. Morrison H, McNally H, Wylie C, McFaul P, Thompson W. The passing score in the objective structured clinical examination. Med Educ 1996 Sept;30(5):345-8.

72. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997.

73. Friedman CP, Wyatt JC. Evaluation methods in medical informatics. New York, NY:Springer, 1997.

74. Ibid., 71.

75. Goroll AH, Morrison G. Core medicine clerkship curriculum guide. Washington, DC: Clerkship Directors in Internal Medicine, 1995.

76. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL, op. cit.

77. Ketchell D. Basic MEDLINE. Seattle, WA: State Resource Service, Health Sciences Library and Information Center, University of Washington, 1988, 1989.

78. Feinglos SJ. MEDLINE, a basic guide to searching. Chicago, IL: Medical Library Association, 1985. (MLA Information Series).

79. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL, op. cit., 546.

80. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 145.

81. Litwin MS. How to measure survey reliability and validity. Thousand Oaks, CA: Sage, 1995:21. (The Survey Kit, No. 7)

82. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 14;98.

83. Wassertheil-Smoller, op. cit., 149.

84. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 145;107.

85. Wassertheil-Smoller, op. cit., 149.

86. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 145;138.

87. SPSS Regression models 9.0. Chicago, IL: SPSS, Inc., 1999:69.

88. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 145;163-4.

89. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. 6th ed. New York, NY: Prentice-Hall, 1997; 145;107.

90. Berk RA. A consumer's guide to criterion referenced test reliability. J Educ Meas 1980 Winter;17(4):323-349.